e-Informatica

software engineering journal

2023 volume 17 issue 1

e-Informatica

e-Informatica

software engineering journal

2023 volume 17 issue 1

e-Informatica

Wrocław University
of Science and Technology

Faculty of Information and Communication Technology, Department of Applied Informatics
Wrocław University of Science and Technology,
50-370 Wrocław, Wybrzeże Wyspiańskiego 27, Poland

e-Informatica Software Engineering Journal
www.e-informatyka.pl, DOI: 10.37190/e-inf

Editorial Office Manager: Wojciech Thomas

Typeset by Wojciech Myszka with the LaTeX 2_ε Documentation Preparation System

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
www.oficyna.pwr.edu.pl;
e-mail: oficwyd@pwr.edu.pl; zamawianie.ksiazek@pwr.edu.pl

ISSN 1897-7979

Print and binding: beta-druk, www.betadruk.pl

Editorial Board

Martin Shepperd (Brunel University London, UK)
Rini van Solingen (Drenthe University, The Netherlands)
Miroslaw Staron (IT University of Göteborg, Sweden)
Tomasz Szmuc (AGH University of Science and Technology Kraków, Poland)
Guilherme Horta Travassos (Federal University of Rio de Janeiro, Brazil)
Adam Trendowicz (Fraunhofer IESE, Germany)
Burak Turhan (University of Oulu, Finland)
Rainer Unland (University of Duisburg-Essen, Germany)
Sira Vegas (Polytechnic University of Madrit, Spain)
Corrado Aaron Visaggio (University of Sannio, Italy)
Bartosz Walter (Poznan University of Technology, Poland)
Dietmar Winkler (Technische Universität Wien, Austria)
Bogdan Wiszniewski (Gdańsk University of Technology, Poland)
Krzysztof Wnuk (Blekinge Institute of Technology, Sweden)
Marco Zanoni (University of Milano-Bicocca, Italy)
Jaroslav Zendulka (Brno University of Technology, The Czech Republic)
Krzysztof Zieliński (AGH University of Science and Technology Kraków, Poland)

Contents

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230101, DOI: 10.37190/e-Inf230101

Governance in Ethical and Trustworthy AI Systems: Extension of the ECCOLA Method for AI Ethics Governance Using GARP

Mamia Agbese*, Hanna-Kaisa Alanen*, Jani Antikainen*, Erika Halme*, Hannakaisa Isomäki*, Marianna Jantunen*, Kai-Kristian Kemell*, Rebekah Rousi*, Heidi Vainio-Pekka*, Ville Vakkuri*

*Faculty of Information Technology, The University of Jvaskyla, Finland

mamia.o.agbese@jyu.fi, hanna-kaisa.h-k.alanen@student.jyu.fi, jani.p.antikainen@student.jyu.fi, erika.a.halme@jyu.fi, hannakaisa.isomaki@jyu, marianna.s.p.jantunen@jyu.fi, kai-kristian.o.kemell@jyu.fi, rebekah.rousi@uwasa.fi, heidi.vainiopekka@gmail.com, ville.vakkuri@jyu.fi

Abstract

Background: The continuous development of artificial intelligence (AI) and increasing rate of adoption by software startups calls for governance measures to be implemented at the design and development stages to help mitigate AI governance concerns. Most AI ethical design and development tools mainly rely on AI ethics principles as the primary governance and regulatory instrument for developing ethical AI that inform AI governance. However, AI ethics principles have been identified as insufficient for AI governance due to lack of information robustness, requiring the need for additional governance measures. Adaptive governance has been proposed to combine established governance practices with AI ethics principles for improved information and subsequent AI governance. Our study explores adaptive governance as a means to improve information robustness of AI ethical design and development tools. We combine information governance practices with AI ethics principles using ECCOLA, a tool for ethical AI software development at the early developmental stages.

Aim: How can ECCOLA improve its robustness by adapting it with GARP® IG practices?

Methods: We use ECCOLA as a case study and critically analyze its AI ethics principles with information governance practices of the Generally Accepted Recordkeeping principles (GARP®).

Results: We found that ECCOLA's robustness can be improved by adapting it with Information governance practices of retention and disposal.

Conclusions: We propose an extension of ECCOLA by a new governance theme and card, #21.

Keywords: AI, AI Ethics, Trustworthy AI, AI Governance, Adaptive Governance, ECCOLA

Submitted: 30 Mar. 2022; Revised: 28 Jul. 2022; Accepted: 18 Sep. 2022; Available online: 28 Sep. 2022

1. Introduction

The continuous progress of artificial intelligence (AI) necessitates that AI ethical development tools and method models used in implementing ethics in the engineering of AI improve their robustness in informing AI governance practices [1]. As AI becomes one of the preferred emerging technologies for software startups [2], which utilize its application in diverse critical sectors such as transport, health, retail, and recently in warfare [3, 4], increased calls for effective AI governance practices are on the rise [5]. Current governance practices implemented in AI engineering lack robustness and often lead to inefficiency in fostering AI governance or failed AI governance practices [5]. The incident involving an autonomous Uber vehicle resulting in a pedestrian's loss of life and the associated failure in identifying the source of accountability [6] due to insufficient governance information represents a growing AI governance failure [7]. Inefficiencies of governance measures in the design and development of the AI posed a legal challenge in clearly delineating responsibility or governance between the malfunctioned AI autonomous driving system and the distracted driver [6]. Consequently, raising questions about the efficacy of current AI governance measures [5].

Governance issues in AI or AI governance concerns often arise when humans interact with AI during usage with no clear delineation of responsibilities associated with roles and the associated impact on humanity and society [8]. The current predominant approach to AI governance is the guideline approach based on AI ethical principles [1]. The approach involves AI principles such as the European Union (EU) Ethics Guidelines for Trustworthy AI [9] used as "soft laws" in the design, development, and deployment of AI to facilitate AI governance [1]. It is also the foundational approach used by AI ethical design and development tools such as method cards, model cards, and ethics canvas to implement AI ethics and subsequently AI governance practices in the development of ethical AI [10, 11]. However, Eitel-Porter [12] explains that the governance practices or ethical principles identified in these principles are insufficient for creating ethical AI that can foster effective AI governance. The guideline approach generally provides a foundation for AI governance but is inadequate due to a lack of information robustness in AI ethics principles [1, 7, 12]. Hamon et al. and Taeihagh [1, 13] corroborate by explaining that information robustness in AI ethical development tools for governance requires a solid information base due to constant changes in the AI terrain. Private information technology organizations that experiment more with AI are often ahead of guidelines and principles in terms of information [1]. Suggesting that insufficient information robustness in ethical AI ethical development tools and method models can lead to duplicated efforts with little practical benefits slowing the pace of research [13]. Overall, Taeihagh [1] stresses the urgency for reassessing current traditional approaches to AI governance to determine their potency as the constant speed of change in information threatens to outpace current AI governance measures [1]. Therefore, more robust governance measures are necessary to manage processes and create associated audits for principles enforcement [1, 13] to help mitigate the increasing inefficiency of AI ethics principles [1, 7].

The adaptive governance and hybrid approach (Adaptive governance) is one of the emerging approaches being explored to improve the robustness of AI governance practices [1, 14, 15]. The adaptive governance approach recommends that successful governance practices be emulated and adapted to current AI governance initiatives to improve AI governance overall [1]. This paper explains that successfully utilized governance practices or frameworks used in regulating previous technologies can be adapted to govern or complement

existing principles or governance approaches for new and emerging technologies such as AI [1]. In this way, lessons learned and successful practices can inform or be incorporated into existing practices to help build robust structures that increase governance capacity [16]. This approach also extends to AI ethical design and development tools to aid the implementation of robust governance practices in the developmental phase that can help mitigate AI governance risks [1, 17]. Currently, most approaches to adaptive governance for ethical AI are explored at policy and regulatory levels [1, 14, 16, 18], with scarce empirical assessment of the approach for AI ethical developmental tools and at the design and development stages [19]. Consequently, virtually no practices have been identified in literature.

Hamon [13] explains that AI ethical development tools can attain a robust or resilient information base by being subjected to rigorous evaluations to benchmark areas that have not been considered or fully exploited. This serves as a motivation for us to explore adaptive governance in AI ethical development tools. In a previous study [20], we analyzed an AI ethical developmental card tool, ECCOLA [10], to identify areas of governance vulnerabilities. ECCOLA is a tool used in the ethically aligned design of trustworthy AI [10] and developed using AI ethics principles (guideline approach). One of the findings from the study revealed a vulnerability in the form of Information governance (IG) practices. We, therefore, leverage the study and extend our research to examine how to improve ECCOLA's robustness by improving its IG vulnerability. We aim to critically analyze ECCOLA's ethical practices with IG practices of the Generally Acceptable Recordkeeping principles® (GARP®) to identify key areas where its information robustness can be improved. By so doing, we can extend the tool for improved governance practices and AI governance by adapting it with IG practices from the GARP® principles. We frame our research question as:

RQ: How can ECCOLA improve its robustness by adapting it with GARP® IG practices?

Our work can help similar AI ethical development tools leverage and optimize the approach to improve their information robustness and add to the scarce body of research on adaptive governance in AI ethical developmental tools.

The rest of the paper is structured as follows: Section 2 focuses on the background and related work on the concepts of AI and its associated technologies, AI ethics, AI governance and the governance frameworks employed in the study. Section 3 describes the methodology employed in this study. Section 4 provides the results, and Section 5 elaborates the findings and discusses them. In Section 6, we provide a conclusion for the research and avenues for further works.

2. Background and related work

2.1. Artificial Intelligence

There is no general or standard definition for AI as it constantly evolves. Collins et al. [21] explains that this is not a challenge as most scientific concepts are truly defined upon maturity. They identify a prevalent definition for AI as systems that mimic reasoning functions associated with human characteristics like learning speech and problem-solving [21]. AI is considered relatively new even though its origin can be traced back to the 1950s with Alan Turing, the Turing machine and Turing's research on making computers

intelligent and capable of replicating the human brain [22]. Paper [23] describes it as various technologies that produce computation associated with human intelligence. AI can be represented as software and possibly hardware systems that act in digital or physical dimensions to perceive their environment, collect data, process it into information, and learn to decide the best course of action to complex goals [24]. AI has evolved over the years to include expert systems (ES) which mirror human intelligence by using knowledge-based applications and inference procedures [25]. ML systems use data to learn from experience with respect to some class of tasks and performance measures [26]. ES includes computer programs or models capable of solving problems in a specific area of knowledge with as much expertise as a human expert [27]. They also automate tasks carried out by human specialists in a particular problem domain. Over time, the growth of the internet alongside computing technology gave rise to the influx of big data, which has evolved AI in a different dimension requiring a ML approach for handling or processing Big data.

ML are computer programs that learn from experience with respect to some class of tasks and performance measures [26]. They employ algorithms in generating numeric models to compute data decisions and are more effective than traditional quantitative methods. ML technology existed separately to AI but has fast become the central paradigm and a sub-field of AI [26]. ML can process structured and unstructured data in real-time to provide accurate predictions and efficiently model predictive data analytic applications and computing tasks where algorithm design is difficult or nearly impossible [26]. ML has evolved over two crucial phases, shallow learning and deep learning (DL) [28]. The shallow learning phase established during the 1990s is characterized by shallow models featuring single or no hidden layers. This phase has seen success in many applications such as web search sorting systems, spam partial filtering systems, and recommendation systems [28]. The DL phase emerged in the 2000s and stems from the study of Artificial Neural Networks (ANN) and how it imitates the human brain's neural structure [28]. The approach follows how the human brain processes information by establishing a simple model, forming different multi-layer learning models with multiple hidden layers and an extensive training data set or sets. Bengio in [29] explains that automatically learning features at multiple extraction levels enables DL systems to learn complex functions and improves their capability to map input functions to output directly from data instead of depending entirely on crafted human features. DL has become well-suited for complex unconstrained problems such as speech and image recognition, and its versatility is harnessed in autonomous and semi-autonomous AI systems [30].

AI is classified into three categories, narrow, general and super intelligence [31]. The narrow stage or phase which is the current state of AI is classified as narrow intelligence. Narrow intelligence is characterised by human-level intelligence (text, speech, and sound = data) to produce outputs such as voice and text recognition capabilities [31]. The other two stages of AI, generalised intelligence and super intelligence which is outside the scope of this study suggests stages where AI systems denote strong human intelligence and above human intelligence respectively. This however, is yet to be achieved [31]. Due to progress made so far in AI, it is increasingly implemented in application areas such as automation of workflow processes, improved service quality, and faster information processing. However, larger data sets used in DL tend to make the network topology complex and challenging to interpret in the design of AI and AI enabled systems. This is one of the leading ethical concerns of AI [30, 32].

2.2. AI ethics

Ethics to begin with is an extremely broad field, crossing diverse disciplines, while possessing roots in normative ethics, which examines what makes actions right or wrong [33]. It involves guidelines or sets of rules and principles to help determine what is good or right and the moral obligations and responsibilities of the entities involved [33]. These entities can range from humans to artificial agents. As artificial agents such as AI become more advanced and their interactions with human agents less defined, concerns have arisen regarding their design and development and the ethical impact of their actions. Some include ethical or moral judgment and decision-making of AI systems, cybersecurity, threats from AI, and goal alignment between humans and AI [8] fuelling the need for ethics to be applied to AI. AI ethics can be described as the field associated with studying ethical issues in AI [33]. The research on AI ethics continues to grow. Michael et al. [34] discuss using ML to design utility systems such as biometrics and how associated biases can be alleviated. Article [35] analyze AI using a contrarian approach on how its dark side, i.e., aspects of AI discussed as negatives, can help influence the development of ethical AI. Their work helps to draw insight into the challenging areas of AI and how these negative aspects can help inform the work on ethical or responsible AI. Trocin et al. [36] helps provide insight into responsible or ethical AI development. They analyze core AI and ML issues from the literature mainly concerned with establishing ethical principles and human values to reduce biases and promote fairness [36]. While their work primarily addresses the healthcare sector, the implications of their findings are relevant to mainstream AI ethics literature.

AI ethics also covers the ethical design and development of AI and the ethical issues that result from AI's interaction with humans [33]. While some have attributed the root of ethical AI issues to design and development in the form of the black-box nature of AI systems [37] others attribute it to their interaction with humans [38]. With all these arguments, the general consensus is for appropriate ethical regulatory measures that can enhance governance practices to be implemented in AI to help address these concerns. One of the main responses to these demands has come in the form of principles or guidelines to act as "soft laws" in regulating AI [1]. Currently, over 80 AI ethics guidelines exist. These stem from different governments, international bodies, private sectors, and the research community [39]. The guidelines or principles center on implementing ethical practices in AI to help mitigate the associated risks. While these principles have helped to identify what is needed to develop ethical AI, they are yet to provide practical solutions on how this will be achieved [40]. As a result, most of the work on AI ethics principles as guidelines focus on frameworks and checklists and not enough on how the guidelines can be implemented as groundwork for governance and innovation [41]. Most sets of principles possess strong similarities to one another, and thus converge in many ways. However, the inconsistencies in how they are interpreted or defined have made it challenging to assign accountability to actors and created room for "digital ethical shopping" [41]. Increased incidence of AI-related accidents continues to occur without appropriate accountability, necessitating that tangible action is needed to move AI ethics from high-level abstractions and arguments to the creation of practical accountability mechanisms [41].

2.2.1. AI governance

In AI ethics, governance has emerged as instrumental in addressing accountability issues where an AI mishap could have monumental consequences [5]. Governance designates how

actions are designed, maintained, regulated and usually represents how accountability is assigned [42, 43]. Governance is a broad concept and rooted in different issues that imply diverse meanings, but generally include processes that facilitate regulation [42, 43]. AI governance or incorporating governance in AI is considered complex and requires several approaches [32]. AI incorporates various technologies such as ML, which is considered to be unpredictable, complex, and random; also, various actors are involved in AI, leading to several conceptualizations of AI governance without an overarching definition [32]. Ashok et al. [44] explain AI governance or the principle of governance in AI as creating and implementing policies procedures and standards for the appropriate development, use, and management of the infosphere, implying that the governance of AI pertains to both the cyberspace and the analog world.

For the implementation of ethics in AI, Sondergaard in [42] explains that AI governance should be explainable, transparent, and ethical, although the meaning of each term is contextual. For example, the term transparency used within technical or legal contexts implies different meanings. Transparency in technology might indicate software or code transparency and in legal context may imply transparency of policies [42]. Therefore, AI governance involves considering several factors and components to aid the governing process. It should also encompass how humanity attempts to navigate the transition of AI as it evolves across different touchpoints [45, 46]. AI governance is considered in research as a layered structure overall which embraces a flexible approach to accommodate the various layers involved [45]. This paper explains that each layer deals with the different ethical and socio-ethical issues associated with AI. For example, the technical layer of AI governance can deal with technical components like data, information, and information security; the political layer can deal with political elements such as regulations, standardization, and the responsible actors [45]. The field of AI governance is still in its infancy [32]. Its current stage is argued as being disorganized, involving different stakeholders vying for their own best interests [47]. Most of the research is centered on different frameworks to aid AI governance at different layers [47], with each tailored to suit a particular context without a clear consensus or framework of how on implementation [5, 45, 46]. This general lack of agreement could be attributed to the tension involved in unifying the different components of AI governance.

One of the dominant and popular approaches to AI governance involves using the principles approach or ethical principles as guidelines for regulation [48–51]. The reasoning is that ethical guidelines can serve as a foundation for regulating AI, encompassing its development, deployment, and usage (its life cycle) [52]. The principles approach encourages the use of guidelines to serve as an ethical risk assessment to help reduce the impact of ethical exposure [46]. However, the principles approach is criticized as ineffective [7, 53]. One contention is that the principles approach lacks methods and practices to translate principles to practice with robust legal and accountability mechanisms to actualize the principles approach lacking [53]. In addition, guidelines may not always be adhered to as they may lack uniformity in the enforcement or be influenced by stakeholders [47]. Further research lends credence to this school of thought by arguing that the principles approach is expensive in terms of costs and processes required in implementation [7].

As a result, there have been arguments for different approaches to AI governance. Some of these approaches include the independent audit governance approach, governance of AI systems in their design, and Adaptive governance [1, 7, 54, 55]. Governance by independent audit calls for an AAA audit style approach to governance [7]. The AAA AI governance approach involves a prospective risk **A**ssessment which entails assessing AI systems before

implementation, and **A**udit trail for failure analysis and accountability; and a system **A**dherence to jurisdictional requirements [7]. Governance of AI systems in their design advocates for governance measures to be applied to AI systems in their design phases to aid accountability and audit in the systems as they evolve [56]. However, the issue of translating guidelines to codes in design is proving to be one of the challenges of enforcing this approach, as ethical guidelines are challenging to translate into codes [1].

The Adaptive Governance approach is the most advocated approach to emerge from the discourse for better AI governance [1]. The adaptive governance approach advocates the co-regulation of governing practices by relevant stakeholders [1]. X. Cao et al. in [57] explains that the concept of adaptive governance goes far back to the 1960s to Arnold Kaufman, who proposed the idea of "participatory democracy." Participatory in the sense that various stakeholders are involved in governance or management processes. Adaptive governance is analyzed as a flexible approach to AI governance [45], where new information is gathered from reiterative adjustment, and guidelines enhancement from successful frameworks [1]. Adaptive governance is also referred to or likened to co-regulation, or hybrid governance approach [15]. With the similarity explained as when AI governance practices are adapted with successful existing practices, a new form of governance emerges, a hybrid of the two [58]. A hybrid approach of governance can help provide flexibility in positioning governance guardrails that proactively identify foreseeable risks emerging from AI as it evolves [1]. Ayling and Chapman [59] corroborates this by explaining that long established techniques exist that help lay down governance practices and having such practices incorporated in the governance of AI can aid AI governance.

Research carried out by [57] identified the adaptive governance approach in a survey as having a significant impact on responsible innovation. Also, Tan et al. [16] proposed an adaptive governance approach as a possible solution for policy regulation of autonomous vehicles in Singapore. However, clear implementation of this approach is yet to emerge in research, particularly at the development and design level of ethical AI systems. Consequently, virtually no practices for adaptive governance as it pertains to AI ethical development tools like ECCOLA have been identified in literature.

2.3. Frameworks

The frameworks used in this study are ECCOLA and the GARP®. ECCOLA represents the principles approach to AI governance and GARP® represents a successful governance frame work.

2.3.1. ECCOLA

The ECCOLA method is an actionable tool to aid the design and development of trustworthy and ethical AI systems presented in a previous study [10]. ECCOLA is considered a low-threshold framework that assists practitioners in their ethically aligned design of AI systems and forms part of the software development process. It also serves as an actionable tool that facilitates AI ethics in a method agnostic manner for AI developers. The method is card-based and comprises 21 cards split into eight ethical themes. The themes are built on ethical principles incorporated from major ethical guidelines, including the IEEE EAD and the EU Trustworthy AI guidelines [10]. The eight themes are analyze, Transparency, Data, Agency and Oversight, Safety and Security, Fairness, Well-being, and Accountability. Each theme comprises one to six cards, and each card provides a detailed approach to the

theme it represents. ECCOLA works by asking AI developers questions to consider and weigh the various ethical issues present in the development of ethical AI systems. It is also method agnostic and can aid software developers, product managers, and consultants with practices and tools for implementing ethically aligned designs in developing AI systems. The questions raised in the cards are based on ethical principles. These are pertinent from the conception stage where ideas are conceived, the mental stage where thinking tools, practices, and principles are analyzed, and the operational stage of the development process. By doing so, ECCOLA embodies the principles approach as the ethical principles on which the method is founded [5], also serves as a governance base in the development process. Table 1 gives a breakdown of the ECCOLA card themes and how they are broken down.

Table 1. ECCOLA card themes

Card themes (8)	Card number (0–20)	Card amount (total 21)
Analze	0	1
Transparency	#1–6	6
Safety and Security	#7–9	3
Fairness	#10–11	2
Data	#12–13	2
Agency and Oversight	#14–15	2
Wellbeing	#16–17	2
Accountability	#18–20	3

2.3.2. GARP® governance framework

Governance frameworks can be described as governance structures that mirror interconnected relationships, factors, and influences within an institution [60]. They typically comprise a conceptual layout with sets of rules on managing and controlling the asset they represent to perform at an efficient level [60]. While many governance frameworks such as information governance (IG), data governance, and corporate governance exist, the focus of our study is on IG to address the gap identified in ECCOLA.

Information governance (IG) is analyzed as a framework that contains mechanisms for guiding the creation, collection, storage, analysis, use distribution, and deletion of information relevant to the business to achieve value creation [61]. IG has its roots in the 20th century when the need arose to develop comprehensive and effective management of increasing volumes of data and information [62]. Previous efforts at governing information focused on basically archiving and retrieving information systematically [62]. However, as the volume of information and particularly electronically stored information increased alongside the speedy development of complex and interlinked systems, it gave rise to policies and rules to govern information and safeguard against loss and distortion [62]. At the time, only large and governmental organizations were inclined to invest in any IG due to the level of complexity, and pricing [62]. With the growth and development of computers and the internet, IG has become a concept available for all institutions, from government and large organizations to small firms and start-ups. In recent times, as the nature of data and information is evolving to include big data and other forms of unstructured data, there is a need to standardize IG and its infrastructure with generally suitable methods and measures [62].

Several IG frameworks exist as IG employs frameworks to enable it to carry out its governance practices [63]. The Unified Governance model also called the Information

Governance Reference model IGRM from the eDiscovery community, EDRM [64] has been developed to address governance issues pertinent to eDiscovery issues. IBM developed an IG model, Information Lifecycle Management which initially focused on data but has since expanded to include other areas of record and information management [64]. The Generally Accepted Recordkeeping Principles (GARP®) or "The Principles" by The Association of Record Managers and Administrators (ARMA) to help address governance of information and records management [65] and even the Control Objectives for Information and Related Technology (COBIT) which focuses on IT governance has been discovered to share some commonalities with GARP® on some non-IT governance requirements such as protection [64]. However, the GARP® framework by ARMA has been recognized as having a widely leveraged global standard which identifies critical hallmarks of IG good practices at a high-level [65]. GARP® is also versatile and agnostic in its approach and can be used within different context [65]. For example [66] used the GARP® in their analysis of IG practices in block chain technology and the General data protection regulation (GDPR), implying that it can be applied to AI development tools such as the ECCOLA and ideal for our study.

GARP® approach provides guidance on information management, and the governance of record creation, organization, security maintenance, and other activities used to support record-keeping [65] efficiently. Records refer to content that could be data or information contextually created, recorded, or received in the initiation, conduct, and completion of an organizational or individual activity [67]. It also encompasses how it relates to other records, the organization or entity that created it, and the metadata likely to define its context. GARP® comprises eight principles – **Accountability**, **Transparency**, **Integrity**, **Protection**, **Compliance**, **Availability**, **Retention**, **Disposition**, and a maturity model made up of five milestones (sub-standard, in development, essential, proactive, and transformational) [65]. However, the scope of this study requires the analysis of GARP® and not the maturity model. The principles are explained further.

1. Accountability: Requires that a senior executive oversees IG programs and delegates responsibilities accordingly for information management, policies, and procedure adoption that guide personnel and ensure auditability.
2. Transparency: This deals with how documentation of activities and processes in an open and verifiable manner is made available to appropriate personnel and interested parties.
3. Integrity: Deals with how IG programs are constructed to reflect the authenticity and reliability of information assets generated or managed.
4. Protection: Deals with IG programs constructed with the appropriate level of protection for information assets on privacy, confidentiality, privilege, secrecy, classified documents, and how they relate to the continuity of the business and protection.
5. Compliance: Deals with how IG programs should be constructed to comply with applicable laws, binding authorities, and organization policies.
6. Availability: Focuses on how IG is exercised in information assets in a timely, efficient, and accurate retrieval manner.
7. Retention: Concerns with how IG is exercised consistently with an organization maintaining its information assets for an appropriate period considering its legal, regulatory, fiscal, operational, and historical requirements.
8. Disposition: Deals with how IG is carried out to provide secure and appropriate disposal of information assets that are no longer required to be maintained in Compliance with organizational laws and policies ([65]).

3. Methodology

Due to the novelty of our study and the limited resources identified in literature, we use ECCOLA as a case study. Case studies help to improve understanding of data derived within a specific context. Case studies within a research is often criticized as being difficult to conduct as they can lead to increased documentation [68]. However, a case study can allow for in-depth interpretation and evaluation of data by aiding the development of conceptual categories to help us add our judgment to the phenomenon found in the data [68]. In addition, since the study involves extending ECCOLA, we also incorporate aspects of design science in conceptualizing the extension of ECCOLA.

3.1. ECCOLA as a case study

Yin [68] explains that a case study can be used to describe or illustrate certain topics within an evaluation in a descriptive manner to provide an in-depth understanding pertinent to the phenomenon under study [68]. Using a case study also provides new or unexpected insights into the subject that can help propose practical courses of action to resolve identified issues and open up possible new directions for future research [68]. Case study is described as a linear but iterative process covering six main steps of planning, designing, preparing, collecting, analyzing and sharing [68]. We explain these steps in the context of ECCOLA.

The planning step aids the researchers in deciding on the use of a case study. Yin [68] explains that case studies are preferred when a "how" question needs to be answered in a research dealing with contemporary issues within real life contexts where control is not the focus. He also explains that case studies are unique in their ability to deal with a full variety of evidence such as documents, artifacts and observations [68]. We explore ECCOLA as a case study based on our research goal hinging on "how" AI ethical tools like ECCOLA can improve their robustness, AI ethical issues transcending technical boundaries to become socio-technical issues within contemporary real-life contexts, and the need for a representative AI ethical developmental tool like ECCOLA to help us deal with the variety of evidence encountered during the course of the study. Yin [68] also explains that a case study approach is for analytical generalization and not for statistical representation which is the case for our study.

For the design process, Yin [68] recommends that case studies are planned in a way that the evidence addresses the research question. As our study focuses on improving information robustness of AI ethical development tools, using such a tool in the form of ECCOLA as our unit of analysis can help ensure that we match the findings to the research question. Also Yin [68] explains that a case can be an event or an entity, of which ECCOLA qualifies as one. As such, we use ECCOLA as a single case study because a single case study can help provide credible test similar to critical experiments [68].

For preparation, Yin [68] explains the need for in-depth preparation to precede the case study. We carried out a comprehensive and rigorous analysis of ECCOLA in our previous study [20] to enable us determine its validity which enabled it for this study. In addition, The GARP® IG framework was carefully sourced from literature as a pertinent source of information practices for the study.

For data collection, [68] explains collection of data or sources of evidence from documentation, archival records, interviews, direct observation, participant observation, illustrative materials and physical artifacts. ECCOLA served as the form of data collection in the form of documentation [68]. Documentation is described as communicable materials that are

used in describing, explaining or instructing regarding attributes of procedures, objects or systems provided using different mediums either digital or analogue [67]. Also, ECCOLA is developed from AI ethics principles from which the method was created to ensure uniformity or convergence. This is because AI ethics principles usually form the basis for most AI ethical development tools like ECCOLA. As such ECCOLA as a data collection for the study also served as the unit of analysis.

Yin [68] explains that the analysis process usually begins with the data collection in case studies. He explains that establishing the data analysis early with well defined analytical tools can help to accomplish many other important aspects of the study [68]. We applied this in our research, as explained earlier, our analysis started with ECCOLA as part of the data collection. The analysis continued using the rigour of content analysis for a more critical analysis with the practices in the GARP® IG framework.

For sharing or communicating the results, Yin [68] recommends that awareness of the audience is important in disseminating the results. Indicating that care must be taken to ensure that the findings are appropriately communicated to the target audience. This is what we aim to do in this study.

3.2. Design science for extending ECCOLA

For the extension of ECCOLA to conceptualize and artifact, we incorporate aspects of the Design Science approach. Hevner et al. [69] explains that design science is naturally a problem solving process which requires knowledge and understanding of the design problem for its solution domain. They outline design processes build and evaluate, where purposeful solutions are built to address hitherto unsolved problems [69]. Our study focuses on the build phase which forms the first guideline out of the seven outlined by [69] for Design Science. They explain that while seven guidelines exists, researchers are not expected to follow them in a mandatory or mechanical fashion but use their creative skills to determine where, when and how to apply the guidelines in each specific projects.

We therefore follow the first guideline, *Design as an artifact* where the research must produce an artifact [69]. However, they explain that artifacts at this stage are rarely full grown information systems used in practice but are ideas and practices through which analysis, design, implementation and use of information systems can be effectively carried out [69]. Arguments exist that building artifacts using design science can be challenging due to the complexities of creative advances in fields with limited theories [69]. However, Design science provides a pragmatic approach that helps extend human and organizational boundaries to create new and innovative artefacts [69], making it suitable for the study. In addition, the pragmatism of the DSRM helps to provide a better process for answering "how" research questions and enables liberalism in exploring the answers [70].

We incorporate aspects from the Design Science research methodology (DSRM) outlined by [71]. Peffers et al. [71] describe the Design science approach as a rigorous process that aims for the design of an artifact to help solve identified problems, make research contributions, evaluate designs and communicate results to appropriate audience [71]. Six process elements or guidelines that can be used to actualize an output is outlined. They include problem identification and motivation, defining the objectives for a solution, Design and Development, Demonstration, Evaluation and communication [71]. However, they explain that while the DSRM processes can follow a nominally sequential order, researchers are not expected to follow the steps rigidly and can start at any step in the process and proceed accordingly [71].

Peffers et al. explains that a problem centered approach can be the basis for the nominal sequence if the research idea results from an observed problem or from suggested future research from a prior project [71]. Our first study [20] provides this entry point. The paper [20] helped us to identify AI governance research gap in the ECCOLA method in the form of IG with the findings communicated in the paper. We therefore continue the process in this paper by following the next two steps as identified by [71]. The processes are defining the objectives of a solution and designing and developing an artifact where we communicate our findings in this paper [71].

3.2.1. Defining the objectives of the solution

For the second stage, the objective of a solution, i.e., what a better artefact can accomplish, we use a literature review to outline benefits of IG practices and how it affects ethical AI systems and development tools such as ECCOLA. The findings are provided to enable a broader understanding of the subject matter and create knowledge on some of the key concepts of IG practices as it pertains to AI governance and AI ethical development tools like ECCOLA [72].

3.2.2. Importance of information governance (IG)

Peffers et al. [71] explains that this stage involves inferring the importance of a solution from the identification of a problem and knowledge of what is possible and feasible. We outline the possibility and feasibility on what is possible with IG practices in AI ethical development tools like ECCOLA.

IG deals with the activities and technologies employed to maximize the value of information and minimize associated risks, and costs [73]. It comprises IG programs and measures that ensure information is appropriately controlled and accessible without compromise [67]. IG is often confused with data governance; however, data governance is narrowly focused on a specific information resource, data which is information in its raw, unstructured and unprocessed form. Data governance mainly deals with how data is governed by implementing appropriate measures and systems for producing and maintaining high-quality data [67]. IG covers information management which deals with managing information assets (IA). Overall, IG strategically provides a framework for managing information legally and ethically and helps balance risks associated with information and the value the information provides [63].

Software developers may find that through utilizing actionable ethical development tools such as ECCOLA, risks associated with the lack of IG governance practices are reduced [63]. Some examples of these risks include storage, disposition, and pre-processing [63]. Improper storage without the guidance of IG can lead to inaccessibility of information and inability to retrieve information. This results in a decrease in value and financial loss for developers. With the current surge in data and Big data used and generated by AI systems during their development life-cycle, having IG guidance on how to store information can help developers avoid this loss. Information appropriately stored in line with IG practices can lead to effective retrieval, and reduced losses from e-discovery and other legal issues [63].

Disposition of information poses another risk for developers of AI items. Disposition risk involves storing too much retrievable data and the risks that may emanate from such practice [63]. Over storing data and information can lead to increased costs and risk using outdated or irrelevant, misleading or ineffective data [63]. Lack of disposition can also

lead to control risk associated with storing data over a prolonged period [63]. When AI systems store data beyond its relevance period, it can pose a risk if hackers gain control of information that should have been long disposed of, putting users' personal information at risk. In addition, if users discover that their personal information supplied to a system at some point is retained after a prolonged period, it can also lead to a lack of trust and confidence in the system. IG advocates for timely disposition of data and information assets so they are not accessed and used maliciously and reduce redundancy costs. For pre-processing risks, IG can aid developers in managing data so that inconsistencies and missing aspects of data that can arise when gathering data are mitigated timely and do not lead to redundancies in the AI systems [63].

3.2.3. Design and development

The third and final step in the nominal process used in this study involves designing and creating an artifact [71]. Peffers et al. [71] explains that artifacts could be potentially models, methods, constructs, instantiations or new properties which could be technical, social or informational resources. They clarify further that a design research artifact could conceptually be any designed object where the results or contributions of a research are embedded in [71]. In the next sections, we present the analysis and findings used in the creation of artifact.

4. Results

This section presents the analysis and presents the findings from the analysis. For the analysis of the study, we employ the use of content analysis. Content analysis enables replicable and valid inferences from texts to the context of their use [74]. It is also useful in evaluating work to compare communication content against previously documented objectives [74]. In addition, content analysis is effective in analyzing text or data such as principles, interviews, field research notes, journals, books, guidelines and reports [75] making it the most suitable option for our study. Using content analysis allowed us to evaluate the languages used within our data, search for bias and make inferences [76]. We also found it useful in reducing our data to concepts to describe the research study by creating categories. Arguments exist that content analysis can be subjective and reductive, but they are also transparent, provide flexibility, and are replicable [76].

4.1. Analysis of ECCOLA with GARP®

We use interpretive content analysis which is qualitative in nature requiring no statistical inference but rather focuses on summarizing and describing meanings in an interpretive and narrative manner [74]. The interpretive approach is described by [74] as a procedure that enables researchers make inference about source and receiver communication from evidence in the messages they exchange. The approach also allows for both manifest and latent content to be considered and analysed [74]. Latent content refers to meaning that is not overt or obvious but is implicit or implied in the communication while manifest content refers to the more obvious meaning within the communication [74].

4.1.1. Process

The content for analysis are the ECCOLA cards. The GARP® framework is used to create an index for the analysis. The index of analysis is created by first defining the units and categories of analysis. The unit of meaning to be coded are identified as activities or processes indicative of the GARP® principles, and the set of categories used for the coding are identified in the table of index, Table 2. A set of rules to determine coding of the ECCOLA cards against the index table are identified as Exist, partially exist and does not exist. The existence of all the practices identified in the index against the card is coded as **Exist**, the existence of one or more but not all is coded as **Partially exist** and the complete absence of activities is coded as **Does not exist**. Each ECCOLA card was manually coded by identifying the units of meaning into the conceptually defined categories in the table of index and the codes documented accordingly [74].

Table 2. Table of index

Unit of meaning	Set of categories
Accountability: activities or processes	Accountability structure Documentation Guiding (policies, procedures, decisions) Audit
Transparency: activities or processes	Open documentation of IA Available documentation of IA Verifiable documentation of IA Accessibility of IA
Integrity: activities or processes	Authentic management of IA, Reliable management of IA
Protection: activities or processes	Protection of private IA Protection of confidential IA Protection of privileged IA Protection of secret IA Protection of essential IA Categorization of IA (private, confidential, privileged, secret, classified)
Compliance: activities or processes	Compliance with applicable laws Compliance with binding authorities Compliance with organizational policies Compliance in (documentation, storage)
Availability: activities or processes	Maintenance of IA for timely retrieval Maintenance of IA for efficient retrieval Maintenance of IA for accurate retrieval Documentation of IA for accessibility
Retention: activities or processes	Maintenance period of IA for legal requirements Maintenance period of IA for regulatory requirements Maintenance period of IA for fiscal requirements Maintenance period of IA for operational requirements Maintenance period of IA for historical requirements Documentation of IA Retention period of IA Storage of IA
Disposition: activities/processes	Secure disposition of irrelevant IA by laws Secure disposition of irrelevant IA by policies Appropriate disposition of IA by laws Appropriate disposition of IA by policies Disposition documentation of IA

4.2. Findings

The result of the analysis is presented in the heat map in Figure 1 and the findings highlighted as Primary Empirical Contributions (PECs).

1. Accountability: The analysis of the ECCOLA cards against the GARP® index of Accountability reveals that four cards (4, 9, 18, and 20) have an "exist" status. Activities and practices within the four cards indicate GARP® IG practices such as transparent

documentation, audit, and activities that demonstrate a **reporting structure** (who makes decisions) or allude to an **accountability structure** aligned with **regulatory bodies** and **policies** [65]. The remaining 17 cards have a status of partially exist. Actions and practices in the cards indicate one or more but not all the IG practices in line with the GARP® index, thus having a partial representation, Indicating that these cards can be improved with GARP® accountability practices. This leads us to our first PEC.

PEC1: 17 of the ECCOLA cards can be adapted fully with GARP® IG practices of Accountability by referencing activities and practices such as documentation, accountability, and auditing of information assets.

2. *Transparency*: The transparency analysis revealed five cards (4, 5, 6, 9, and 19) with an "exist" status showing clear indications of practices and activities that facilitate **documentation** and **accessibility** of IA in a clear and verifiable manner **accessible** by the appropriate personnel. The remaining 16 cards have a "partially exist" status. There are indications of either of the practices in these cards, but not all of them, For example, the practices and activities in card #8 (Data quality) point towards proper accessibility of data but makes no reference to documentation or mode of documentation of IA that can result from these practices forming PEC2.

   PEC2: ECCOLA can be adapted fully with The GARP® IG Transparency practices with references such as documentation, mode of documentation, and accessibility in 16 cards.
3. *Integrity*: The integrity analysis indicates that all the 21 ECCOLA cards reflect activities and practices that align with Integrity practices of the GARP®. Activities and practices within the cards indicate **authenticity** and **reliability** in the handling of IA affiliated with the GARP® of Integrity. In Card #10 (Human Agency), the activities and practices are geared at sensitizing developers on the need for authentic practices that lead to reliability for the AI users. While some cards do not state the practices expressly, there is an allusion to them leading us to the third PEC.

   PEC3: ECCOLA has exist status and does not require further adaptation with the GARP® principle of Integrity.
4. *Protection*: The Protection analysis reveals eight cards (6, 7, 8, 9, 12, 13, 18, and 20) have an "exist" status. GARP® IG practices such as **protection mechanisms** for **designated** or **categorized** IA exist in these cards. One of the cards #7 (Privacy and Data) sensitizes developers on the need for protection of data by asking questions on encryption, anonymization of data, and accessibility of protected IA. In the remaining 13 cards, either one of these practices was identified, resulting in a "partially exist" status. This forms the fourth PEC.

   PEC4: The ECCOLA method can be adapted fully with GARP® practices of Protection with references such as protection/protection mechanisms and categorization or designation of IA in 13 cards.
5. *Compliance*: The compliance analysis reveals that 13 cards, (1, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, and 20) have an "exist" status by displaying IG practices that align with the GARP® index. Implying that these cards create awareness for AI developers of compliance practices such as **documentation, storage, applicable laws, organizational policies**, and **authorities** for IA. Only one or two of the practices could be identified in the remaining eight cards forming our PEC5.

   PEC5: ECCOLA can be adapted fully with Compliance practices of the GARP®

principle in eight cards by making references such as documentation, storage, applicable laws, organizational policies, and authorities for IA.

6. *Availability*: The Availability analysis shows that only one card has an "exist" status. Activities and processes within the lone card #9 ask AI developers questions such as who has access to users' data and the circumstances they are granted access – ensuring that IA **documentation**, **accessibility**, and **retrieval** align with the Availability GARP® IG practices in the index. The remaining 20 cards have a "partially exist" status by indicating one or two of these practices but not all of them.
   PEC6: ECCOLA can be adapted fully with the GARP® practices of Availability in 20 of its cards by referencing practices of documentation, accessibility, and retrieval of IA.
7. *Retention*: The Retention analysis reveals that nine of the cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) have a "partially exist" status. The Cards asks AI developers questions on GARP® Retention IG practices such as **documentation, storage**, **maintenance** and **retention/period** in one or two contexts or allude to them in activities for IA. However, the remaining 12 cards show no indication or allusion to these practices leading us to our PEC7.
   PEC7: ECCOLA has partially exist status of GARP® principle of Retention in nine cards and does not exist status in 12 cards. ECCOLA can be extended to incorporate the GARP® IG practices of Retention such as documentation, storage, maintenance and retention.
8. *Disposition*: The Disposition analysis reveals that nine of the cards have a "partially exist" status. The cards (0, 4, 5, 6, 7, 8, 9, 18, and 20) have one or two GARP® Disposition practices and activities such as **documentation, transfer**, and **Disposition** of IA that can facilitate IG. In comparison, these practices could not be identified in the remaining 12 cards. Card #3 (Communication) asks AI developers pertinent questions on the need for transparent practices in developing AI systems; however, there are no questions or practices on how generated IA are disposed of, leading us to PEC8.
   PEC8: ECCOLA has partially exist status of GARP® principle of Disposition in nine cards and does not exist in 12 cards. ECCOLA can be extended to incorporate GARP® IG practices of Disposition such as documentation,transfer, and Disposition of IA.

5. Discussion

We discuss the outcome of the analysis which yielded eight PECs in Table 3. The PECs and their implications are discussed in this section.

PEC1 is based on adapting accountability practices of documentation, accountability, and auditing of information assets to ethical AI development tools like ECCOLA to help facilitate audit practices. The absence of some of its practices can translate to crucial accountability practices being omitted while using the ECCOLA cards. Where these practices are not present, software developers may cultivate practices and activities that overlook accountability structure or non-documentation of crucial information that does not comply with policies. In addition, adapting GARP® practices with already existing ethical principles can help ECCOLA mitigate accountability risks that may arise in the development of AI systems [1].

Reddy et al. [5] analyses accountability as a challenge as regards implementation in terms of governance. They explain that appropriate stages are needed for effective accountability practices and stress the need for an approval structure by governing bodies or regulating

ECCOLA CARD	GARP® by ARMA							
	Accountability	Transparency	Integrity	Protection	Compliance	Availability	Retention	Disposition
#0 Stakeholder analysis								
#1 Types of transparency								
#2 Explainability								
#3 Communication								
#4 Documenting trade-offs								
#5 Traceability								
#6 System Reliability								
#7 Privacy and Data								
#8 Data quality								
#9 Access to data								
#10 Human agency								
#11 Human oversight								
#12 System security								
#13 System safety								
#14 Accessibility								
#15 Stakeholder participation								
#16 Environmental Impact								
#17 Societal Effects								
#18 Auditability								
#19 Ability to redress								
#20 Minimizing negative impacts								
	Partially exist	Does not exist	Exist					

Figure 1. ECCOLA analysis with GARP®

authorities that oversee and preview processes to ensure proper documentation of IA that can aid audits in governance [5]. This article further explains that accountability in governance requires regulatory oversight and needs to be in place in the development stage of AI. Guiding actions by an accountability structure and explanations in the form of IA documentation can facilitate internal or external audits. It also makes developers of AI accountable in the documentation of their IA and may help reduce the opacity of AI governance [5]. The principle of transparency (PEC2) highlights the need for transparent practices such as documentation, mode of documentation of IA to make them more accessible in IG for AI developers. Caron [77] argues that transparent processes in AI systems help to improve auditability. She explains that open and verifiable practices that generate IA must be transparent in the development process. Adapting transparency governance practices can aid understanding when IA is accessed by appropriate personnel, which is vital for auditability. Also, in developing AI, different cognitive biases and heuristics exist [77] warranting the need for transparent obligations to be imposed. These obligations can be in the form of auditable governance measures to help mitigate these practices so that when these IA are accessed, there is a clear understanding that helps audit in governance [77]. Kiener in [37] agrees with this argument and discusses the need for open and verifiable processes in the development of AI in sensitive fields like medicine. He explains that transparent processes and activities of AI can aid human oversight, risks, and audits in governance frameworks like IG.

PEC3 highlights the principle of integrity, which explains the need for reliability and authenticity in the processes, activities, and practices that generate IA in AI development. Adapted practices of integrity help facilitate proper audits in governance frameworks. Jansen et al. [78] explains that a governance framework in the development of AI can enable authentic and reliable IA. They analyze that a governance structure (such as IG) can enable IA of integrity by managing the quality, validity, security, and associated risks in practice to preserve the integrity.

PEC4 is based on the principle of protection. It emphasizes the need for protection practices such as protection/protection mechanisms and categorization or designation of IA for designated IA (private, confidential, privileged, secret, classified). Protection mechanisms and practices can provide safeguards to mitigate risks from incidental access or disclosures. Culnan in [79] supports this and explains that the IG protection practices in the development of AI can help ensure they are secure and properly categorized to help avert security and privacy breaches. While introducing regulatory mechanisms such as the GDPR and the California Consumer Privacy Act (CCPA) exists to help protect AI users, they can be insufficient in protecting IA that are not adequately categorized or labeled. To ensure that a suitable form of protection is provided [79], incorporating these practices can aid the audit processes involved in governance for AI developers [79].

PEC5 findings relate to how compliance practices like documentation, storage of IA in a manner that complies with applicable laws, organizational policies, and other binding authorities adapted in the development stage of AI can help aid AI governance overall. In et al. [80] argues that IG practices of maintaining IA in a manner that conforms to compliance (internal and external) can help mitigate risk and increase efficiency. He explains that when developers or organizations familiarise themselves with compliance practices and streamline the maintenance of their IA in line with governance frameworks like IG, such practices become routine and make it easy to produce systems compliant with applicable laws and binding authorities. Furthermore, in legal matters and regulation, these practices can help audit processes in AI governance [80].

PEC6 is based on the principle of availability. It explores how adapting practices such as documentation, accessibility, and retrieval practices can facilitate the maintenance of IA to ensure timely efficiency. Hind et al. [81] explains that AI developers are usually faced with the challenge of documentation of IA as there are no clear guidelines on how much to document to provide enough clarity. Therefore, governance practices geared towards appropriate documentation, accessibility, and retrieval of IA to ensure that IA is effectively and adequately documented and, upon retrieval, provide holistic information may aid clarity. Documentation of IA in line with a governance framework like IG also provides confidence that information made available is wholesome and suitable for all interested parties. In addition, these practices also aid the governance frameworks in audit processes to ensure unity [81].

PEC7 is based on the principle of retention and how incorporating governance practices such as effective maintenance, documentation, storage, and retention (period of retention) of IA can improve AI governance in development methods like ECCOLA. Kroll in [82] recommends the minimization of retention of collected records or the disposal of aggregate records where possible to enhance efficient governance of records. He explains that retention of IA should be appropriately maintained, documented and subject to a governance structure to reduce the risks of retaining them beyond their retention period. Retention of IA combined with the principles approach can help minimize the risk from legitimate requests from law enforcement [82]. When Information Assets are retained beyond their life cycle, they can pose a risk if authorities request them and utilize them beyond the scope for which they were acquired [82].

PEC8 works on how adapting disposition practices like secure and appropriate transfer, disposition, and documentation of records or information (IA) in compliance with applicable laws and policies can improve AI governance in development methods like ECCOLA. Kroll [82] explains that regular disposal of aggregate and redundant records or reducing them to the lowest level of sensitivity can reduce privacy risks and increase the efficiency of AI governance. When Information Assets are maintained for a period, a need for them must

be further established to enable them not to pose a risk of redundancy which may hamper efficiency. A clear need exists for records or information to be retained for a period and disposed of accordingly, as it can help make development methods more trustworthy when AI audits are carried out.

Table 3. Primary Empirical Contributions (PECs)

Primary Empirical Contributions (PECs)
PEC1 – 17 of the ECCOLA cards can be adapted fully with GARP® IG practices of Accountability by referencing activities and practices such as documentation, accountability, and auditing of information assets.
PEC2 – ECCOLA can be adapted fully with GARP® IG Transparency practices with references such as documentation, mode of documentation, and accessibility in 16 cards.
PEC3 – ECCOLA has exist status and does not require further adaptation with the GARP® principle of Integrity.
PEC4 – The ECCOLA method can be adapted fully with GARP® practices of Protection with references such as protection/protection mechanisms and categorization or designation of IA in 13 cards.
PEC5 – ECCOLA can be adapted fully with Compliance practices of the GARP® principle in eight cards by making references such as documentation, storage, applicable laws, organizational policies, and authorities for IA.
PEC6 – ECCOLA can be adapted fully with the GARP® practices of Availability in 20 of its cards by referencing practices of documentation, accessibility, and retrieval of IA.
PEC7 – ECCOLA has partially exist status of GARP® principle of Retention in nine cards and does not exist status in 12 cards. ECCOLA can be extended to incorporate the GARP® IG practices of Retention such as documentation, storage, and retention.
PEC8 – ECCOLA has partially exist status of GARP® principle of Disposition in nine cards and does not exist in 12 cards. ECCOLA can be extended to incorporate GARP® IG practices of Disposition such as documentation, transfer, and Disposition of IA.

5.1. Extension of ECCOLA

From the discussion, it has been identified that ECCOLA can be improved and extended to incorporate IG practices to improve its information robustness. While six of the principles can be improved in the cards by making references to pertinent practices, a need exist for proper representation of the unidentified principles of retention and disposition and governance practices as a whole. The discussion has also outlined how the lack of governance principles identified and particularly those of retention and disposition can pose risks for developers of AI systems who use tools like ECCOLA. As such, it is important to highlight these key governance practices in the form of a new theme and card. As such, we propose a new governance theme in ECCOLA and a new card which incorporates the deficient IG practices of retention and disposition. Therefore, we propose a new card, #21 under a new theme, governance illustrated in Figure 2.

Motivation Letting users know the period of retention of their records by the AI system creates Transparency and encourages trust that their records will not be stored indefinitely.

What to do: Ask yourself:

– Are records maintained in line with policies?
– Are users informed of the retention period of their records?
– Are users informed of the need to retain their records beyond the retention period?

Figure 2. Card 21

– Are obsolete and redundant records destroyed or disposed in line with regulatory requirement?
– Are records securely disposed and not retained beyond their retention period?

Practical example: Establishing clear time limits for collected records and explaining the need to retain records beyond retention period makes the system trustworthy. AI systems that provide recruitment recommendations will find their services ineffective if users' information is not updated regularly to reflect their skill set.

When AI developers create a culture of openly communicating to users and end-users, the period of retention and disposition of their records or information, it can enable them to trust such AI or AI-enabled systems. The trust can be due to increased confidence that their information or records will not be stored and used indefinitely [82]. Also, frequently updating users' records and disposing of redundant records can enable developers of AI systems ensure that they have access to current and better quality records or information [82].

5.1.1. Significance of a new theme and card

To a large extent the discourse on governing information assets as it relates to AI ethics has largely focused on data governance. This may explain the major focus of proposed ethical model tools such as google model cards, data ethics canvas and even the current version of ECCOLA on data governance. These tools help draw attention of AI developers and relevant stakeholders to data ethics governance challenges which are vital in implementing AI ethics. The current version of ECCOLA has visible representation of data governance practices in the data cards(#7, #8 and #9) but there is no clearly delineated representation for IG practices for records or information that can be generated from data.

Model cards as AI ethical development tools help to provide *information* on trained machine learning models [83] or AI models to improve ethical practices such as transparency.

As such it is pertinent that data governance practices are in place for data collected or sourced at the basic data level. But the records generated from the data used in training these models which are actually represented as *information* require that IG practices are in place for these sets of information assets as well. As Glasser et al. [45] explains that AI governance is a layered structure requiring governance practices at each layer to achieve the goal of governance. Therefore, incorporating IG practices such as retention and disposal and highlighting them in a governance theme can improve improve information robustness and help create a culture of visibility and informed communication for both developers and users [45].

Therefore, in comparison to other proposed ethical AI development tools that promote AI governance practices like the current version of ECCOLA, the proposed theme and card can help provide visibility and improve the information base of these tools towards AI governance. The current ECCOLA has no clear governance representation as most of them are embedded in the AI ethics practices. However the growing emphasis on the need for visible governance practices to be implemented at the developmental stages [1, 5, 10, 39, 45] necessitates that AI ethical developmental tools like ECCOLA have visible measures that aim to improve governance information and to tackle governance challenges that can occur in development. Visibility of Governance measures at this stage can help bring to the forefront pertinent AI governance challenges and provide the necessary information needed to address them [5, 45].

5.2. Validity threats

The reliability of the content analysis is a potential validity threat to the research. While interpretive content analysis is reproducible [74], the study could be subject to our own interpretation due to the nature of subjectivity of the documents such as the AI ethical principles in the ECCOLA cards and the researchers in determining the index terms from the GARP® documents used in the actual analysis.

The use of a single case study also poses a validity threat to the study. As highlighted by [68], the use of a single case study can pose a challenge for generalization of results. However, [68] still explains that a single case study can be used for analytic generalization and not for statistical generalization which is the case for our study.

A third and possibly one of the biggest threats to the validity of our study is the incomplete Design science approach used. We understand that most Design science usually require a valid artifact as output for the study [69]. Following the build and and evaluate approach [69], our study at this stage focuses on the build stage or the design an artifact stage. We are aware that this can serve as a limitation to the study, however, we argue that following the recommendations of [69], we are not following all the methodology steps in a rigid or mechanical fashion and will use the outcome of each study to enable us leverage the next study until we have gone through all the DSRM steps to produce a viable artifact [69].

6. Conclusions and further works

This study explored how ethical practices in AI ethical developmental tools like ECCOLA can be adapted with IG practices for improved information robustness in tackling AI governance issues using the adaptive or hybrid governance highlighted by [1]. ECCOLA

was critically analysed with the GARP® IG framework using content analysis to highlight areas where the practices in the cards could be improved with IG practices for a more information robust base. The results reveal 8 PECs, which indicate that most of the IG practices in the GARP®: Accountability, Transparency, Integrity, Protection, Compliance, and Availability are represented in ECCOLA either partially or fully to varying degrees. The findings further reveal that all 21 cards in ECCOLA comply with IG practices of Integrity. However, the principles of Retention and Disposition were shown to be the least represented and lacking IG practices in ECCOLA as they could not be identified in 12 of the cards. Implying that ECCOLA can improve its information robustness in all 21 cards and also be extended with a governance theme to improve its governance practices.

Regarding the implication of the findings to AI governance, it may imply that AI ethical practices in ECCOLA may be insufficient in addressing some governance issues that may arise in the development process of AI systems. AI governance issues that deal with Integrity may be fully addressed and partially addressed for Accountability, Transparency, Protection, Compliance, and Availability practices. However, issues that may arise from Retention and Disposition AI governance challenges may be partially or not addressed, indicating that the method's ethical practices can be improved in terms of its information robustness to aid AI governance. As a solution, the index terms from the analysis are suggested as potential modifiers for the cards with partial practices. However, for a better representation of governance practices in the tool, a governance theme is proposed. The theme will help address pertinent governance issues for developers as well as improve governance information robustness of ECCOLA. The first of the cards in this theme is proposed as #21 which addresses concerns from governance issues as it pertains to retention and disposition. The card provides motivation, suggested activities, and a practical example of how the card can effectively tackle ethical AI governance challenges that may arise at the developmental stage. Therefore, the analysis offers an answer and a possible solution to our research question.

In addition to analyzing ECCOLA, this study also explored an adaptive AI governance approach initiative at the development stage. The ethical practices were modified or adapted with IG practices to create a more practical approach to governance issues that developers of AI systems can encounter at the development stage.

The findings or outcome from this study will form the basis for the next phase of our study which will involve evaluation of the artifact to continue the DSRM process. Paper [71] explains that the DSRM is not a linear process to be followed rigidly but allows for the incorporation of different elements of the methodology as the research progresses. Therefore, we aim to continue the remaining parts of the research in subsequent studies which are outside the scope of this one.

Information for funding/support or the lack of it

This research is partially funded by Business Finland (business-finland.fi)research projects. The are Sea4value and Stroke Data and ITEA4. The authors are grateful to the funders for their support.

References

[1] A. Taeihagh, "Governance of artificial intelligence," *Policy and Society*, 2021, pp. 1–21.

[2] M. Schulte-Althoff, D. Fürstenau, and G.M. Lee, "A scaling perspective on AI startups," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021, p. 6515.
[3] C. Giardino, S.S. Bajwa, X. Wang, and P. Abrahamsson, "Key challenges in early-stage software startups," in *International conference on agile software development*. Springer, 2015, pp. 52–63.
[4] C. Newton, J. Singleton, C. Copland, S. Kitchen, and J. Hudack, "Scalability in modeling and simulation systems for multi-agent, AI, and machine learning applications," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, Vol. 11746. International Society for Optics and Photonics, 2021, p. 1174626.
[5] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, "A governance model for the application of AI in health care," *Journal of the American Medical Informatics Association: JAMIA*, Vol. 27, No. 3, 2020, pp. 491–497.
[6] P. Liu, M. Du, and T. Li, "Psychological consequences of legal responsibility misattribution associated with automated vehicles," *Ethics and Information Technology*, Vol. 23, No. 4, 2021, pp. 763–776.
[7] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura et al., "Governing AI safety through independent audits," *Nature Machine Intelligence*, Vol. 3, No. 7, 2021, pp. 566–571.
[8] S. Du and C. Xie, "Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities," *Journal of Business Research*, Vol. 129, 2021, pp. 961–974.
[9] "Ethics guidelines for trustworthy AI," European Commission, High-Level Expert Group on AI, Tech. Rep., 2019. [Online]. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[10] V. Vakkuri, K.K. Kemell, M. Jantunen, E. Halme, and P. Abrahamsson, "ECCOLA – A method for implementing ethically aligned AI systems," *Journal of Systems and Software*, Vol. 182, 2021, p. 111067.
[11] D. Lewis, W. Reijers, H. Pandit, and W. Reijers, *Ethics canvas manual*, ADAPT Centre and Trinity College Dublin and Dublin City University, 2017. [Online]. https://www.ethicscanvas.org/download/handbook.pdf
[12] R. Eitel-Porter, "Beyond the promise: Implementing ethical AI," *AI and Ethics*, Vol. 1, No. 1, 2021, pp. 73–80.
[13] R. Hamon, H. Junklewitz, and I. Sanchez, "Robustness and explainability of artificial intelligence," 2020.
[14] I. Linkov, B.D. Trump, K. Poinsatte-Jones, and M.V. Florin, "Governance strategies for a sustainable digital world," *Sustainability*, Vol. 10, No. 2, 2018, p. 440.
[15] U. Pagallo, P. Aurucci, P. Casanovas, R. Chatila, P. Chazerand et al., "On good AI governance: 14 priority actions, a SMART model of governance, and a regulatory toolbox," 2019.
[16] S.Y. Tan and A. Taeihagh, "Adaptive governance of autonomous vehicles: Accelerating the adoption of disruptive technologies in Singapore," *Government Information Quarterly*, Vol. 38, No. 2, 2021, p. 101546.
[17] S. Jain, M. Luthra, S. Sharma, and M. Fatima, "Trustworthiness of artificial intelligence," in *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020, pp. 907–912.
[18] I. Brass and J.H. Sowell, "Adaptive governance for the Internet of Things: Coping with emerging security risks," *Regulation and Governance*, Vol. 15, No. 4, 2021, pp. 1092–1110.
[19] A.B. Whitford and D. Anderson, "Governance landscapes for emerging technologies: The case of cryptocurrencies," *Regulation and Governance*, Vol. 15, No. 4, 2021.
[20] M. Agbese, H.K. Alanen, J. Antikainen, E. Halme, H. Isomäki et al., "Governance of ethical and trustworthy AI systems: Research gaps in the ECCOLA method," in *29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2021, pp. 224–229.
[21] C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, "Artificial intelligence in information systems research: A systematic literature review and research agenda," *International Journal of Information Management*, Vol. 60, 2021, p. 102383.
[22] A.B. Simmons and S.G. Chappell, "Artificial intelligence-definition and practice," *IEEE Journal of Oceanic Engineering*, Vol. 13, No. 2, 1988, pp. 14–42.

[23] S. Leijnen, H. Aldewereld, R. van Belkom, R. Bijvank, and R. Ossewaarde, "An agile framework for trustworthy AI," in *Proceedings of the First International Workshop on New Foundations for Human-Centered AI*, 2020, pp. 75–78.
[24] "A definition of AI: Main capabilities and scientific disciplines," European Commission, High-Level Expert Group on AI, Tech. Rep., 2019. [Online]. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
[25] C.F. Tan, L. Wahidin, S. Khalil, N. Tamaldin, J. Hu et al., "The application of expert system: A review of research and applications," *ARPN Journal of Engineering and Applied Sciences*, Vol. 11, No. 4, 2016, pp. 2448–2453.
[26] L. Ma and B. Sun, "Machine learning and AI in marketing–Connecting computing power to human insights," *International Journal of Research in Marketing*, Vol. 37, No. 3, 2020, pp. 481–504.
[27] E.J. Rykiel Jr., "Artificial intelligence and expert systems in ecology and natural resource management," *Ecological Modelling*, Vol. 46, No. 1–2, 1989, pp. 3–8.
[28] P. Hu, Y. Lu et al., "Dual humanness and trust in conversational AI: A person-centered approach," *Computers in Human Behavior*, Vol. 119, 2021, p. 106727.
[29] Y. Bengio, *Learning deep architectures for AI*. Now Publishers, Inc., 2009.
[30] R.R. Kumar, M.B. Reddy, and P. Praveen, "Text classification performance analysis on machine learning," *International Journal of Advanced Science and Technology*, Vol. 28, No. 20, 2019, pp. 691–697.
[31] K. Oosthuizen, E. Botha, J. Robertson, and M. Montecchi, "Artificial intelligence in retail: The AI-enabled value chain," *Australasian Marketing Journal*, No. 3, 2020.
[32] B.W. Wirtz, J.C. Weyerer, and B.J. Sturm, "The dark sides of artificial intelligence: An integrated AI governance framework for public administration," *International Journal of Public Administration*, Vol. 43, No. 9, 2020, pp. 818–829.
[33] K. Siau and W. Wang, "Artificial intelligence (AI) ethics: Ethics of AI and ethical AI," *Journal of Database Management (JDM)*, Vol. 31, No. 2, 2020, pp. 74–87.
[34] K. Michael, R. Abbas, P. Jayashree, R.J. Bandara, and A. Aloudat, "Biometrics and AI bias," *IEEE Transactions on Technology and Society*, Vol. 3, No. 1, 2022, pp. 2–8.
[35] P. Mikalef, K. Conboy, J.E. Lundström, and A. Popovič, "Thinking responsibly about responsible AI and 'the dark side' of AI," *European Journal of Information Systems*, Vol. 31, No. 3, 2022, pp. 257–268.
[36] C. Trocin, P. Mikalef, Z. Papamitsiou, and K. Conboy, "Responsible AI for digital health: A synthesis and a research agenda," *Information Systems Frontiers*, 2021, pp. 1–19.
[37] M. Kiener, "Artificial intelligence in medicine and the disclosure of risks," *AI and Society*, Vol. 36, No. 3, 2021, pp. 705–713.
[38] V.C. Müller, "Ethics of artificial intelligence and robotics," in *The Stanford Encyclopedia of Philosophy*, E.N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. [Online]. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/
[39] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, Vol. 1, No. 9, 2019, pp. 389–399. [Online]. http://www.nature.com/articles/s42256-019-0088-2
[40] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices," in *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, 2021, pp. 153–183.
[41] M. Hickok, "Lessons learned from AI ethics principles for future actions," *AI and Ethics*, Vol. 1, No. 1, 2021, pp. 41–47.
[42] P. Sondergaard, *AI governance – what are the KPIs? And who is accountable?*, 2021.AI, (2019, Nov). [Online]. https://2021.ai/ai-governance-kpi [Accessed: 16 Jun. 2021].
[43] R.I. Rotberg, "Good governance means performance and results," *Governance*, Vol. 27, No. 3, 2014, pp. 511–518.
[44] M. Ashok, R. Madan, A. Joha, and U. Sivarajah, "Ethical framework for artificial intelligence and digital technologies," *International Journal of Information Management*, Vol. 62, 2022, p. 102433.

[45] U. Gasser and V.A. Almeida, "A layered model for AI governance," *IEEE Internet Computing*, Vol. 21, No. 6, 2017, pp. 58–62.
[46] A.F. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical AI and autonomous systems," *Proceedings of the IEEE*, Vol. 107, No. 3, 2019, pp. 509–517.
[47] J. Butcher and I. Beridze, "What is the state of artificial intelligence governance globally?" *The RUSI Journal*, Vol. 164, No. 5-6, 2019, pp. 88–96.
[48] H. Yu, Z. Shen, C. Miao, C. Leung, V.R. Lesser et al., "Building ethics into artificial intelligence," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18. AAAI Press, 2018, p. 5527–5533.
[49] A. Daly, T. Hagendorff, H. Li, M. Mann, V. Marda et al., "AI, governance and ethics: global perspectives," The Chinese University of Hong Kong, Faculty of Law, Research Paper 2020/05, 2020. [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3684406
[50] B. Perry and R. Uuk, "AI governance and the policymaking process: Key considerations for reducing AI risk," *Big data and cognitive computing*, Vol. 3, No. 2, 2019, p. 26.
[51] W. Wu, T. Huang, and K. Gong, "Ethical principles and governance technology development of AI in China," *Engineering*, Vol. 6, No. 3, 2020, pp. 302–309.
[52] P. Cihon, "Standards for AI governance: International standards to enable global coordination in AI research and development," Future of Humanity Institute, University of Oxford, Technical Report, 2019. [Online]. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf
[53] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, Vol. 1, No. 11, 2019, pp. 501–507.
[54] R. Leenes and F. Lucivero, "Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design," *Law, Innovation and Technology*, Vol. 6, No. 2, 2014, pp. 193–220.
[55] M. Firlej and A. Taeihagh, "Regulating human control over autonomous systems," *Regulation and Governance*, Vol. 15, No. 4, 2021, pp. 1071–1091.
[56] K. Yeung, A. Howes, and G. Pogrebna, "AI governance by human rights-centered design, deliberation, and oversight," in *The Oxford Handbook of Ethics of AI*. Oxford University Press, 2020, p. 77.
[57] X. Cao, D. Lv, L. Zhang, and Z. Xing, "Adaptive governance, loose coupling, forward-looking strategies and responsible innovation," *IEEE Access*, Vol. 8, 2020, pp. 228 163–228 177.
[58] R. Radu, "Steering the governance of artificial intelligence: National strategies in perspective," *Policy and Society*, 2021, pp. 1–16.
[59] J. Ayling and A. Chapman, "Putting AI ethics to work: Are the tools fit for purpose?" *AI and Ethics*, Vol. 2, 2021, pp. 405–429.
[60] O.E. Williamson, "The economics of governance: Framework and implications," *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics*, 1984, pp. 195–223.
[61] H. Borgman, H. Heier, B. Bahli, and T. Boekamp, "Dotting the I and crossing (out) the T in IT governance: New challenges for information governance," in *49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2016, pp. 4901–4909.
[62] H. Dogiparthi, "History of information governance," University of the Cumberlands, Dept. of Information Technology, Research Paper, 2019. [Online]. https://www.researchgate.net/publication/330844911_History_of_Information_Governance
[63] E.M. Coyne, J.G. Coyne, and K.B. Walker, "Big data information governance by accountants," *International Journal of Accounting and Information Management*, 2018.
[64] J. Hagmann, "Information governance–beyond the buzz," *Records Management Journal*, 2013.
[65] *Generally Accepted Recordkeeping Principles®*, ARMA International, 2017. [Online]. https://www.arma.org/page/principles
[66] D. Hofman, V.L. Lemieux, A. Joo, and D.A. Batista, ""The margin between the edge of the world and infinite possibility": Blockchain, GDPR and information governance," *Records Management Journal*, 2019, pp. 240–257.

[67] E. Lomas, "Information governance: information security and access within a uk context," *Records Management Journal*, No. 2, 2010.
[68] R.K. Yin, *Case Study Research: Design and Methods*. Sage Publications, 1994.
[69] A.R. Hevner, S.T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, No. 1, 2004, pp. 75–105.
[70] J.W. Creswell, W.E. Hanson, V.L. Clark Plano, and A. Morales, "Qualitative research designs: Selection and implementation," *The Counseling Psychologist*, Vol. 35, No. 2, 2007, pp. 236–264.
[71] K. Peffers, T. Tuunanen, M.A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, Vol. 24, No. 3, 2007, pp. 45–77.
[72] R.E. Stake, *The art of case study research*. Sage Publications, 1995.
[73] S. Bennett, "What is information governance and how does it differ from data governance?" *Governance Directions*, Vol. 69, No. 8, 2017, pp. 462–467.
[74] J.W. Drisko and T. Maschi, *Content analysis*. Pocket Guide to Social Work Re, 2016.
[75] R.P. Weber, *Basic content analysis*, Quantitative Applications in the Social Sciences. Sage, 1990, No. 49.
[76] S. Elo, M. Kääriäinen, O. Kanste, T. Pölkki, K. Utriainen et al., "Qualitative content analysis: A focus on trustworthiness," *SAGE Open*, Vol. 4, No. 1, 2014.
[77] M.S. Caron, "The transformative effect of AI on the banking industry," *Banking and Finance Law Review*, Vol. 34, No. 2, 2019, pp. 169–214.
[78] M. Janssen, P. Brous, E. Estevez, L.S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy artificial intelligence," *Government Information Quarterly*, Vol. 37, No. 3, 2020, p. 101493.
[79] M.J. Culnan, "Policy to avoid a privacy disaster," *Journal of the Association for Information Systems*, Vol. 20, No. 6, 2019, p. 1.
[80] J. In, R. Bradley, B.C. Bichescu, and C.W. Autry, "Supply chain information governance: Toward a conceptual framework," *The International Journal of Logistics Management*, 2018.
[81] M. Hind, S. Houde, J. Martino, A. Mojsilovic, D. Piorkowski et al., "Experiences with improving the transparency of AI models and services," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.
[82] J.A. Kroll, "Data science data governance [ai ethics]," *IEEE Security and Privacy*, Vol. 16, No. 6, 2018, pp. 61–70.
[83] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman et al., "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 220–229.

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230102, DOI: 10.37190/e-Inf230102

The Effect of Dual Hyperparameter Optimization on Software Vulnerability Prediction Models

Deepali Bassi*, Hardeep Singh*
*Department of Computer Science, Guru Nanak Dev University, India
deepalics.rsh@gndu.ac.in, hardeep.dcse@gndu.ac.in

Abstract

Background: Prediction of software vulnerabilities is a major concern in the field of software security. Many researchers have worked to construct various software vulnerability prediction (SVP) models. The emerging machine learning domain aids in building effective SVP models. The employment of data balancing/resampling techniques and optimal hyperparameters can upgrade their performance. Previous research studies have shown the impact of hyperparameter optimization (HPO) on machine learning algorithms and data balancing techniques.
Aim: The current study aims to analyze the impact of dual hyperparameter optimization on metrics-based SVP models.
Method: This paper has proposed the methodology using the python framework Optuna that optimizes the hyperparameters for both machine learners and data balancing techniques. For the experimentation purpose, we have compared six combinations of five machine learners and five resampling techniques considering default parameters and optimized hyperparameters.
Results: Additionally, the Wilcoxon signed-rank test with the Bonferroni correction method was implied, and observed that dual HPO performs better than HPO on learners and HPO on data balancers. Furthermore, the paper has assessed the impact of data complexity measures and concludes that HPO does not improve the performance of those datasets that exhibit high overlap.
Conclusion: The experimental analysis unveils that dual HPO is 64% effective in enhancing the productivity of SVP models.

Keywords: software vulnerability, hyperparameter optimization, machine learning algorithm, data balancing techniques, data complexity measures

1. Introduction

With the advent of information technology, Software is the main component of devices and systems on which modern life is dependent. Software Vulnerabilities are mistakes, errors, flaws, weaknesses, or loopholes caused during the specification, design, development, or configuration of the software [1]. We can say that a lack of programming practices [2] may lead to software vulnerability which may further provide a gateway for attacks, thereby hampering the confidentiality, integrity, and availability of the information systems.

Submitted: 30 Jun. 2022; Revised: 8 Oct. 2022; Accepted: 18 Nov. 2022; Available online: 5 Dec. 2022

Attackers make use of software vulnerabilities to attack the system. Once access is obtained, security is at stake and any valuable information can be stolen from it, for example, theft of email passwords, debit card information, etc., or the system can be corrupted. To curb the intrusion of attackers, software free from vulnerabilities needs to be developed using security techniques along with secured design principles and software development life cycles [3]. Conventional security techniques include static analysis [4] such as penetration testing [5], fuzz-testing [6], etc., dynamic analysis such as tainted data-flow analysis [7], code inspections [8], etc., and hybrid analysis [9]. Static analysis has the problem of a high false-positive rate [10]. Code reviews are time-consuming and cannot be easily performed on large systems because of time constraints, and scarcity of validation and verification resources. To tackle these limitations to an extent, researchers worked on developing new prediction models based on machine learning algorithms. The time of the developers is reduced by early prediction of the most vulnerable components.

Predictive modeling calculates or predicts future outcomes using historical data and aids in prioritizing the efforts of software testing. The prediction models consist of independent variables (predictors) and dependent variables (outcomes). They are built after collecting historical data for relevant predictors. In the domain of software engineering, various fault prediction and defect prediction models are proposed for predicting faults and defects and enhancing the efficiency of software testing plans [11]. Therefore, to predict vulnerabilities researchers developed SVP models. Although faults and vulnerabilities are different, yet construction of their respective models is almost similar [12]. SVP machine learning-based models classify the software components into vulnerable or non-vulnerable categories depending on different levels of granularity such as file, package, class, or method. They are categorized as metrics-based, text-mining-based, and a combination of both. Metrics-based models are where metrics determine the vulnerable components. Text-mining-based models are where the conversion of source code into tokens and frequencies predicts the vulnerable components. A combination of both predicts the vulnerability by combining the metrics and text features.

Vulnerabilities are hard to find, as it requires attack patterns knowledge and understanding of the source code. Due to security concerns, developers publish a limited number of vulnerabilities compared to faults [13] so, vulnerabilities are subgroups of faults. The class imbalance problem has always been an issue as prediction models favor the majority class (over-represented concept) over the minority class (under-represented concept). Various studies have managed to tackle it by using data-level methods [14–18], algorithm-level methods [19], and ensemble learning methods [20].

1.1. Motivation

The main challenge faced by the SVP models is their performance which is affected by the imbalanced datasets and the hyperparameter settings of machine learning techniques. Therefore, to enhance the efficacy of prediction models, recent studies have used various combinations of resampling techniques and optimized machine learning methods [21–25] in the areas of software engineering. By far, studies have optimized the hyperparameters of machine learners and applied data balancing techniques to balance the dataset but optimization of resamplers has only been explored in a few studies [26–29]. In [26], Six imbalanced datasets from the Keel collection are used. The current study aims to analyze the effect of hyperparameter optimization (HPO) when applied to both machine learners and resamplers called dual HPO on the performance of metrics-based SVP models which

has not been performed on the PHP dataset before. In addition to this, it has been observed that the capability of machine learners is not only hampered by imbalanced datasets but also by the degree of class overlapping, which motivates us to further anatomize the problem with the degree of class overlapping for the cases where efficiency is not improved. Papers [27–29] have worked on bug prediction models and worked on textual data.

1.2. Contributions

In this study, we have replicated the six scenarios of HPO used in [26] on three publicly available datasets PHPMyAdmin, Moodle, and Drupal [30]. We have used five supervised machine learning algorithms and five resampling techniques which are highly used in past studies. This paper revolves around the construction of metrics-based SVP models using

Table 1. Scenarios for hyperparameter optimization

Scenario	Machine Learning Methods	Resampling Techniques
1. Ad + Rn	Default hyperparameters	No Resampling
2. Ao + Rn	Optimized hyperparameters	No Resampling
3. Ad + Rd	Default hyperparameters	Default hyperparameters
4. Ao + Rd	Optimized hyperparameters	Default hyperparameters
5. Ad + Ro	Default hyperparameters	Optimized hyperparameters
6. Ao + Ro	Optimized hyperparameters	Optimized hyperparameters

the six scenarios mentioned in Table 1. The comparative analysis of all the scenarios is performed, and significant improvement is calculated using Wilcoxon signed-rank test, and Bonferroni correction is applied as there are six statistical tests. The effect of hyperparameter tuning and resampling on the prediction models is demonstrated through the outcomes of our experiments performed in the python framework Optuna. The study has the following major contributions:

RQ 1) How much is dual HPO effective in improving the performance of SVP models?
Dual HPO is the optimization of hyperparameters of both machine learners and resampling techniques. Since HPO is known to improve the productivity of SVP models [21–25] the aim is to check whether applying dual HPO increases their performance.

RQ 2) Is dual HPO better than other HPO scenarios?
This question aims to find whether the performance improvements by dual HPO compared with AdRd are better than HPO scenarios AoRd and AdRo when each is compared with AdRd.

RQ 3) How has the degree of class overlapping affected the HPO?
As mentioned in [26] the performance of SVP is also affected by the degree of class overlapping which is measured by the data complexity measures. This paper has used two measures imbalance ratio and maximum Fisher's discriminant ($F1$). The current study is the first attempt to search for the cause of no improvement in the efficiency of models even after applying HPO.

RQ 4) Which resampling technique has performed the best?
The current study has applied five resampling techniques namely SMOTE, Adasyn, Borderline SMOTE, SMOTE + Tomek links, and SMOTE + Edited Nearest Neighbors to balance the datasets. Four scenarios in Table 1 have used resampling with default and optimized parameters respectively. So, the goal of this research question is to find the best resampling technique across all datasets and machine learning techniques.

1.3. Paper organization

The remaining sections of the current study are structured as Section 2 presents the related work, Section 3 provides the methodology, Section 4 describes the experimental design, Section 5 explains the results, Section 6 discusses the findings of the study, Section 7 shows threats to validity, and Section 8 concludes the paper and provides the future scope of the study.

2. Related work

Software vulnerability provides the gateway for attackers to damage the software systems. Therefore, research studies have focused on predicting the software vulnerable components effectively using machine learning algorithms. There exists a large amount of work that exhibits machine-learning techniques for the construction of SVP models. The performance of models is increased by using machine learners but more research needs to be done on the factors affecting them, i.e., hyperparameter tuning, imbalanced datasets, and the degree of class overlapping. In this paper, we are trying to improve machine learners' performance by understanding the role of these factors.

2.1. Class imbalance in prediction models

Ghaffarian and Shahriari [1] have presented a survey paper showing machine learning and data-mining techniques to mitigate the impact of software vulnerability. It has encapsulated various definitions of software vulnerability. Various works related to vulnerability prediction models have been reviewed in this paper. Also, it has been mentioned that the vulnerability datasets are imbalanced and affect the productivity of machine learning algorithms.

Kaya et al. [3] have described the effect of feature types, machine learners, and resamplers on SVP models. It has included three feature types' metrics, text, and a combination of both. It has experimented using seven machine learners namely random forest, linear support vector machine, weighted k-nearest neighbors, adaboost, rusboost, linear discriminant, subspace discriminant and four resamplers such as SMOTE, cluster SMOTE, borderline SMOTE, and adasyn. The datasets used are Drupal, Moodle, and PHPMyAdmin. For experimental evaluation, the performance metrics used are precision, recall, AUC, $F1$-score, and specificity. The paper concludes that random forest has performed the best for smaller datasets Drupal and PHPMyAdmin, and rusboost for larger datasets, i.e., Moodle.

Wang and Yao [15] have used under-sampling techniques, ensemble-learning techniques, and threshold moving techniques on two classifiers (naive bayes and random forest) to address the class imbalance issue in software defect prediction (SDP) models. It has been concluded that balanced random under-sampling shows better defect prediction but lesser than naive bayes. Adaboost has turned out to be the best performer in improving the efficiency of SDP models. The overall performance is assessed using G-mean, AUC, and balance metrics.

Sasada et al. [16] have raised the data complexity issue caused by resampling techniques that affect the accuracy of the predictive model. The performance metrics used are accuracy and f-measure. SMOTE + ENN and SMOTE + TL are proposed for handling the noise and overlap in the dataset. The proposed method is effective in four out of ten datasets. Borowska and Stepaniuk [17] have studied the behavior of under-sampling and over-sampling methods

on imbalanced datasets. The experimental results illustrated that SMOTE + ENN and SMOTE + TL provide good results for datasets with few positive instances and random oversampling (ROS) gives good results when there are more positive instances.

Walden et al. [30] have stated that previous works were done on Java, C++, and C projects so they have proposed the vulnerability datasets based on PHP open-source projects as most of the vulnerabilities are found in web applications. The datasets provided include software metrics and text features. The paper has worked on both software metrics-based and text mining-based SVP models and found that text mining-based models perform better than metrics-based ones. Experiments include random forest machine learner and under-sampling technique for balancing the dataset and the performance metrics used are recall and inspection ratio.

Stuckman et al. [31] have extended the work in [30] by inspecting the impact of dimensionality reduction techniques (feature selection, principal component analysis, and confirmatory factor synthesis) on metrics and text-mining features. It has implemented SMOTE and under-sampling technique for data balancing and found that SMOTE has shown lower recall, lower inspection rate, and higher f-measure so; it is preferred over the under-sampling method. In addition to this, dimensionality reduction techniques worked well for cross-project prediction than within-project prediction.

2.2. Hyperparameter tuning

Hyperparameter tuning in recent studies has been observed to improve the efficiency of prediction models. A lot of studies exist on hyperparameter tuning of defect prediction and bug prediction models which motivated us to explore the same for vulnerability prediction models.

Tantithamthavorn et al. [21] have shown the effect of optimal parameter settings on the defect prediction model (DPM) by using the caret technique (automated parameter optimization technique) and concluded that on optimizing the parameters of classifiers, the performance of DPM has improved by 40% in terms of AUC evaluation metric. The paper has suggested experimenting with different parameter tuning of machine learning classifiers.

Rijn and Hutter [22] have employed a hundred datasets from OpenML to find the important hyperparameters for the random forest, adaboost, and support vector machine algorithms. This paper projects the idea of optimizing important hyperparameters rather than optimizing all the hyperparameters. We have considered some of these important hyperparameters to be tuned as per our research requirement.

Yang and Shami [24] have identified the hyperparameters for machine learning algorithms through their work. The study has discussed the HPO problem in detail and has explained different HPO algorithms and frameworks with their advantages and disadvantages. The range for hyperparameters of (both classifiers and regressors) random forest, k-nearest neighbor, and support vector machine. For our study, we have used classifiers' ranges.

Shu et al. [25] perform parameter tuning of machine learners and data pre-processors to classify bug reports. A comparison of FARSEC and HPO is performed. It is observed that on applying HPO, the recall has improved from 35% to 65% with an increased false-positive rate. Also, optimizing data pre-processors produces better performance results than optimizing machine learners. The paper used five machine learners such as random forest, logistic regression, naïve bayes, k-nearest neighbor, and multilayer perceptron.

Claesen and Moor [32] have discussed the challenges in searching hyperparameters for the machine learning algorithm. Also, current approaches for searching hyperparameters are

mentioned in the paper. Kudjo et al. [33] have discussed the importance of parameter tuning on three PHP datasets Drupal, Moodle, and PHPMyAdmin. The paper has compared the results for random forests with the benchmark study by [30] and found an increase in precision, recall, and accuracy for Drupal and PHPMyAdmin whereas for Moodle only accuracy has increased. The paper has not included balancing techniques. In addition to this, the accuracy metric gives biased results for imbalanced datasets.

Sara et al. [34] use the concept of data balancing and HPO on maintainability or bug prediction models. The paper uses SMOTE and grid search on five machine learning algorithms (k-nearest neighbor, support vector machine, decision tree, naive bayes, and multilayer perceptron). The evaluation metrics involve sensitivity, specificity, accuracy, and precision. Grid search is better than default settings in all datasets but there is no existence of the best machine learning technique for all datasets. However, it concludes that tuning hyperparameters and balanced data helps in obtaining the best productivity of machine learning methods.

Osman et al. [35] have worked on optimizing hyperparameters of two machine learning algorithms (k-nearest neighbor and support vector machine) using five open-source java systems. Hyperparameter tuning improves the accuracy of bug prediction models such as k-nearest neighbor became a better predictive model after HPO. The paper shows how different machine learning algorithms are compared after hyperparameter tuning.

2.3. Class overlapping

Almutairi and Janicki [14] discuss the class imbalance problem and the impact of overlapping on imbalanced data. It has compared and analyzed three machine learning algorithms (decision tree, k-nearest neighbor, and support vector machine) using six combinations of overlapping and imbalance. For the evaluation criteria accuracy, precision, and recall are used. The findings show that imbalanced data has less overlap than balanced data. It further concluded that the performance of the predictive model not only depends on resampling techniques but also on the degree of overlapping in the datasets.

Barella et al. [36] have explained the class overlapping issue in imbalanced binary classification. It states that various research studies have focused on balancing methods but that works well when the classes are linearly separable. The class overlap is measured by various data complexity measures mentioned in Ho and Basu [37] and Sotoca et al. [38] that divide the data complexity measures into three categories: (i) measures of feature overlap, (ii) class separability measures, and (iii) measures of geometry and topology.

Furthermore, Lorena et al. [39] have divided these measures into the following feature-based measures linearity measures, neighborhood measures, network measures, dimensionality measures, and class imbalance measures. Also, it has mentioned the application areas for data complexity measures such as data analysis, data pre-processing tasks, learning algorithms, and meta-learning. For our study, since we need to deal with data imbalance which is one of the data pre-processing tasks, therefore, data complexity measures are applied to see their impact on the classification of imbalanced datasets.

2.4. Dual hyperparameter optimization

Shu et al. [27] extended the paper [25] by incorporating dual optimization approach (SWIFT) for bug prediction. The research paper concludes that SWIFT is better than optimizing data pre-processor and learners individually. Kong et al. [26] show the impact

Table 2. Comparisons with existing works

Research work	Machine learning techniques used	Evaluation metrics used	Resampling	HPO	Dual HPO	Data complexity measures	Datasets
Kaya et al. [3]	RF, AB, Linear Discriminant, Linear SVM, Weighted KNN, Subspace Discriminant, Rusboost	*AUC*, Precision, Recall, *F*-Score, Specificity	SMOTE, Adasyn, ClusterSMOTE, BLSMOTE	NO	NO	NO	PHP
Shu et al. [25]	RF, NB, Logistic Regression, Multilayer Perceptron, *k*-nearest Neighbors	Precision, Recall	SMOTE	YES	NO	NO	Chromium, Apache projects (Ambari, Wicket, Camel, Derby)
Walden et al. [30]	RF	Recall, Inspection Ratio	Under-sampling	NO	NO	NO	PHP
Stuckman et al. [31]	RF	*Recall*, *F*1-Score, Inspection Ratio	Under-sampling, SMOTE	NO	NO	NO	PHP
Kudjo et al. [33]	RF, KNN, SVM, Decision Tree	Precision, Recall, Accuracy	No Resampling	YES	NO	NO	PHP
Zhang et al. [40]	RF, NB, Decision Tree	Recall, Precision, Accuracy	No Resampling	NO	NO	NO	PHP
Abunadi et al. [41]	NB, Logistic Regression, SVM, RF, Decision Tree	Precision, Recall, F-Measure	No Resampling	NO	NO	NO	PHP
Khalid et al. [42]	RF, NB, Decision Tree	*Accuracy*, *Precision*, *Recall*, *F*1-Score	SMOTE	NO	NO	NO	PHP
Catal et al. [43]	Averaged Perceptron, Bayes point machine, Boosted Decision Tree, Decision Forest, Decision jungle, Deep SVM, SVM, Logistic Regression, Multilayer Perceptron	*AUC*	No Resampling	NO	NO	NO	PHP
Shu et al. [27]	RF, NB, Logistic Regression, Multilayer Perceptron, *k*-nearest Neighbors	*Precision*, *Recall*, *F*1-score, False alarms, G-mean	SMOTE	YES	YES	NO	Chromium, Apache projects (Ambari, Wicket, Camel, Derby)
Kong et al. [26]	RF, SVM	*AUC*	SMOTE, Adasyn, SMOTE + Tomek links, SMOTE + Edited Nearest Neighbor	YES	YES	YES (maximum Fisher's discriminant)	Keel-Collection
Proposed Work	RF, AB, NB, KNN, SVM	*AUC*, *F*1-Score	SMOTE, Adasyn, BLSMOTE, SMOTE + ENN, SMOTE + TL	YES	YES	YES (maximum Fisher's discriminant and imbalance ratio)	PHP

of resampling and HPO on the performance of machine learning algorithms. Also, tuning the hyperparameters of both resamplers and learners is emphasized. The area under the ROC curve is used for performance evaluation. The experiments are performed on two machine learning algorithms (random forest and support vector machine) that consider six combinations of HPO (learners and resamplers). Also, data complexity measures show that hyperparameter optimization works for datasets with low overlap than datasets with high overlap. Existing works have encouraged us to propose a methodology that assimilates data imbalance, hyperparameter tuning, and class overlapping areas to increase the efficacy of SVP models.

Agrawal et al. [28] have also applied HPO on pre-processor and machine learners for defect prediction models where the "dodge" technique is applied to reduce the CPU cost caused by HPO. It eliminates duplicate hyperparameter tunings but is efficient for data with low dimensionality; therefore Agrawal et al. [29] extended the work for high-dimensionality data. In [28] 10 SE defect prediction datasets and 6 SE issue tracking datasets were used. In addition to this, [29] has included 63 SE datasets that explore Github issue close time, 4 SE datasets for bad smell detection, and 37 non-SE problems from the UCI repository hence considering datasets with high dimensionality.

2.5. Comparisons with existing works

Table 2 has compared our work with the research studies based on machine learning techniques used, resampling techniques, evaluation metrics, whether HPO exists, dual HPO exists, and whether data complexity measure exists. It has been inferred from Table 2 that most of the studies lack HPO, dual HPO, and data complexity measures. In [27], only data complexity measures are missing. Since [26] has included all the factors but on the Keel-collection datasets, only two machine learning techniques, four resampling techniques, and one data complexity measure. In addition to this, only (Ao + Ro) is compared with (Ad + Rd).

The proposed work has replicated this idea on the PHP dataset with three added machine learning algorithms, i.e., gaussian naïve bayes, adaboost, and k-nearest neighbor, one added resampler namely Borderline SMOTE, and one added data complexity measure such as imbalance ratio. Furthermore, six scenarios are compared and mentioned in Section 4.2.

3. Research methodology

This section explains the research methodology stating the datasets used (Section 3.1), machine learning methods (Section 3.2), resampling techniques (Section 3.3), hyperparameter optimization (Section 3.4), performance evaluation metrics (Section 3.5), and data complexity measures (Section 3.6) used in the study.

3.1. Experimental datasets

The experiments are performed on three open-source publicly available datasets[1], namely Drupal is a content management system, PHPMyAdmin is an open-source administration for MySQL, and Moodleis a learning management system. These datasets have been used

[1]http://seam.cs.umd.edu/webvuldata

in recent studies mentioned in the related work section. The level of granularity of datasets is file and each file is labeled with "no" (no vulnerability exists) or "yes" (at least one vulnerability exists from vulnerability type). There exist vulnerable types such as code injection, cross-script request forgery (CSRF), cross-site scripting (XSS), path disclosure, authorization issues, and others related to phishing or man-in-the-middle vulnerabilities. The current paper focuses on binary classification as the dataset available are labeled with two values "NO" and "YES". There exist both metrics and text mining datasets1. For this work, only metrics-based datasets are considered. The datasets downloaded have comma-separated values which are further preprocessed and saved. Table 3 shows the Drupal vulnerability prediction dataset after preprocessing. Each column header is the software metric of the dataset. There are 13 software metrics (Independent variable) and the column "IsVulnerable" (dependent variable) shows the labeling of each file:

– "nonecholoc": non HTML lines of code;
– "loc": total number of lines of code in a PHP file;
– "nmethods": total number of functions in the file;
– "ccomdeep" and "ccom": cyclomatic complexity, i.e., the number of independent paths. Since these two metrics have the same values so can be considered as one metric;
– "nest": maximum nesting complexity, i.e., maximum depth of the nested loops;

Table 3. Example of Drupal dataset

non echoloc	loc	nmethods	ccom deep	ccom	nest	hvol	nIncoming Calls	nIncoming CallsUniq	nOutgoing InternCalls	nOutgoing ExternFlsCalled	nOutgoingExtern FlsCalledUniq	nOutgoingExtern CallsUniq	Is Vulnerable
4	4	0	1	1	0	3.29	0	0	0	2	2	2	0
126	126	9	26	26	4	1402.86	18	6	2	18	8	35	1
168	168	10	29	29	3	1455.88	4	4	8	19	9	32	0
412	412	35	77	77	5	4929.76	361	119	23	30	12	30	1
10	10	3	1	1	0	29.93	54	33	0	0	0	0	0
53	53	3	16	16	4	583.07	53	32	1	14	6	24	0
1355	1355	92	251	251	8	17945.97	698	147	38	62	19	137	1
162	162	13	31	31	3	1702.981	59	34	4	24	10	16	1
172	172	21	29	29	2	1854.86	113	77	7	17	5	35	0
131	131	19	14	14	1	1455.23	176	79	3	16	6	25	0
135	135	19	14	14	1	1510.92	176	79	3	16	6	25	0
328	328	42	51	51	2	4440.29	286	84	15	27	6	67	0
381	381	27	84	84	6	5256.54	48	15	16	24	9	82	1
896	896	70	197	197	6	13130.98	107	51	27	38	11	108	1
107	107	9	18	18	3	941.85	1	1	3	4	4	14	0
77	77	9	13	13	3	917.13	8	4	3	4	3	4	0
361	361	22	86	86	6	4169.19	33	9	8	26	10	39	0
60	60	2	11	11	1	360.58	0	0	1	2	2	16	0
59	59	2	10	10	1	413.19	0	0	1	2	2	16	0
74	74	2	16	16	1	473.45	0	0	1	2	2	18	0

– "hvol": Halstead's volume is calculated using the number of total operands and operators with the number of unique operators and operands;
– "nIncomingCalls": fan-in, i.e., number of files that call the function from the measured file;
– "nIncomingCallsUniq": internal methods that are called by the statement in the measured file;
– "nOutgoingInternCalls": fan-out is the number of files called from the measured file;
– "nOutgoingExternFlsCalled": total external calls are the number of methods from other files called in the measured file;
– "nOutgoingExternFlsCalledUniq": external methods called are the number of functions called from the measured file and also included in other files;
– "nOutgoingExternCalls": external calls to methods are the number of files that calls methods in the measured file.

Table 4 describes the version of the project, total files, vulnerable files, Imbalance Ratio (IR), no of vulnerabilities, Maximum Fisher's Discriminant Ratio ($F1$), and text features. As per equation (8) mentioned in Section 3.6, IR for each dataset is calculated and it is observed that all three datasets are imbalanced, Moodle being highly imbalanced with IR 120.8. Drupal and PHPMyAdmin are imbalanced with IR 2.25 and 10.92, respectively. Equation (7) in Section 3.6 calculates the $F1$ values of the datasets. Moodle is less overlapped with the $F1$ value of 0.8098. Drupal is moderately overlapped, having a 0.6229 $F1$ value, and PHPMyAdmin is highly overlapped with an $F1$ of 0.3195.

Table 4. Dataset descriptions

Dataset	Version	Total Files	Vulnerable Files	IR	Vulnerabilities	$F1$	Text Features
Drupal	6.0	202	62	2.25	97	0.6229	3886
PHPMyAdmin	3.3.0	322	27	10.92	75	0.3195	5232
Moodle	2.0.0	2924	24	120.8	51	0.8098	18306

3.2. Machine learning methods

Machine learning methods consist of supervised and unsupervised algorithms. Supervised machine learning methods include those machine learning algorithms where input features are mapped to the target using labeled data. These include Linear Regression, Logistic Regression, Naive Bayes (NB), k-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), etc. Various ensemble learning methods are Bagging, Voting, Adaboost, etc. In this paper, five supervised machine learning methods; RF, AB, Gaussian NB, SVM, and KNN are used. In paper [44], we have studied the effect of hyperparameter optimization on eight machine learning algorithms that are widely used in research studies as per Table 2 and have different training strategies. Further, we have considered the top five machine learners with high AUC values and extended our work on those techniques.

3.2.1. Random forest (RF)

Most studies use the RF algorithm as mentioned in Table 2; therefore, it is selected for the current study. It is a supervised learning technique based on the concept of ensemble

learning. By counting the votes (classification output) of multiple decision trees constructed from randomly generated subsets in the forest, the class with the majority votes is selected to be the classification output of the RF classifier. It can perform classification and regression tasks. It can handle large datasets and prevent the over-fitting issue [45].

3.2.2. Adaboost (AB)

Adaboost, called adaptive boosting, is the boosting algorithm to build strong classifiers using several weak classifiers. Weak learners are sequentially added and trained by weighted training data. It predicts the classification output by calculating the weighted mean of the weak classifiers. It boosts the efficiency of the prediction model in binary classification problems. It can use different base learners to improve its performance. Noisy data and outliers highly affect the AB algorithm [46–48].

3.2.3. Naive Bayes (NB)

Naive Bayes is the supervised machine learning algorithm based on Bayes' theorem, assuming there is independence among the features of the class. NB models are of four types: Gaussian NB (GNB), Multinomial NB (MNB), Bernoulli NB (BNB), and Complement NB (CNB) [49] and [50].

- In GNB, the predictors follow the gaussian distribution. MNB is implemented for multinomial distributed data and used for text classification.
- BNB is used for multivariate Bernoulli distributed data where multiple features exist, with each feature to be assumed as binary-valued. CNB is a variant of MNB and is suitable for imbalanced datasets. NB is easy to implement and consumes less time but has the limitation of independence among predictors, which in real-life cases can affect the performance of the classifier.
- We used GNB in this paper as compared to CNB because the former has performed better than the latter [51].

3.2.4. Support vector machine (SVM)

SVM is a supervised machine learning algorithm used to construct a hyperplane (best decision boundary) that separates n-dimensional space into classes to put new data points in the correct category for the future. SVM are linear and non-linear used for linearly separable and non-linearly separable data, respectively. It is effective for high dimensional space, is memory efficient, and consists of various kernel functions (linear, polynomial, radial basis function, and sigmoid) used for decision function. It has the limitation of overfitting, which arises when the number of samples is much smaller than the number of features [52, 53].

3.2.5. k-nearest neighbors (KNN)

KNN is a supervised machine learning algorithm that classifies the data points by calculating the distance between them. It classifies the new data point based on the similarity with the stored data. It is crucial to determine the value of k as a small value may lead to underfitting and a large value to overfitting [54, 55].

3.3. Resampling techniques

Resampling is an approach to yielding a balanced dataset (training dataset) from an imbalanced dataset (training dataset). Resampling can be done in three ways: Under-sampling, Over-sampling, and Hybrid resampling. Under-sampling removes the majority of samples leading to loss of data and degrading the performance of the classifiers. For example, random under-sampling (RUS). Over-sampling replicates the minority samples which leads to over-fitting. It should be ensured that over-sampling is restricted only to the training dataset to avoid over-fitting issues. Examples are: random oversampling (ROS), synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling approach (ADASYN), borderline SMOTE (BL SMOTE), cluster SMOTE, etc. [56]. Hybrid sampling is the combination of under-sampling and over-sampling techniques such as SMOTE + edited nearest neighbor (SMOTE + ENN), SMOTE + Tomek links (SMOTE + TL), and condensed nearest neighbors + Tomek links (CNN + TL).

In this study, the main focus is on over-sampling and hybrid sampling techniques so; we have used SMOTE, ADASYN, BL SMOTE, SMOTE + ENN, and SMOTE + TL for our research.

3.3.1. SMOTE

SMOTE generates synthetic samples by creating new instances rather than duplicating the existing ones. The oversampling process occurs in the feature space. The synthetic samples are yielded along with the line segments that join k-nearest neighbors of the minority samples. It prevents over-fitting and increases the performance capabilities of the classifiers by generalizing the decision boundaries [57].

3.3.2. ADASYN

ADASYN produces synthetic samples by giving importance to minority samples that are hard to learn and minority classes having fewer samples. Density distribution is considered in ADASYN. It promotes adaptive learning and reduces learning bias [58].

3.3.3. BL SMOTE

BL SMOTE is an extended version of SMOTE. In this approach, synthetic samples are generated by selecting misclassified instances of the minority class. The instances near and on the borderline are misclassified than instances far from the borderline [59].

3.3.4. SMOTE + TL

SMOTE + TL is a hybrid sampling technique that integrates SMOTE and Tomek links to reduce the occurrence of overlap. SMOTE oversamples the minority class, and Tomek links are removed from the oversampled samples. A clear decision boundary is formed by removing the instances in the overlapping region [16].

3.3.5. SMOTE + ENN

SMOTE + ENN combines the SMOTE technique with the ENN to reduce the noise. SMOTE oversamples the minority class, and ENN removes the noisy samples. It removes

misclassified examples by its three nearest neighbors. It deeply cleanses the data more than SMOTE + TL [16].

3.4. Hyperparameter optimization

Optimal hyperparameter settings are required to improve the potential of SVP models [33]. The data learning process initializes and updates the parameters, known as model parameters, but we cannot estimate hyperparameters from this process. Hyperparameters are set before the model's training, as they configure prediction models and minimize the loss function. Hyperparameter tuning can either be done: manually or automatically.

In the case of manual tuning, various machine learning algorithms require a deep knowledge of hyperparameters. It is laborious and time-consuming for algorithms with larger hyperparameters. Due to the above limitations, the optimization of hyperparameters is automated called hyperparameter optimization (HPO). It is time-efficient, reduces human effort, aids in comparing machine learning algorithms, and determines the suitable prediction model for a particular problem [32].

There exist various HPO techniques, and selecting the apt technique is crucial. Grid search (GS), random search (RS), bayesian optimization (gaussian process, SMAC, tree-structured Parzen estimator), gradient-based optimization, multi-fidelity optimization algorithms (successive halving, hyperband, bayesian optimization hyperband), and metaheuristic algorithms (genetic algorithm, particle swarm optimization) are the HPO techniques [24]. These are applied depending on the hyperparameters such as continuous, conditional, categorical, and discrete.

Some open-source libraries to handle the HPO problems are sklearn, spearmint, bayesopt, hyperopt, optunity, and optuna [24]. For this paper, Optuna [60] is used instead of hyperopt [26] due to its advantages that search space is dynamically constructed; searching and pruning algorithms are efficient, scalable, lightweight, and distributed.

Different machine learning algorithms have different hyperparameters to be tuned. Search spaces are selected based on recent studies. Table 5 describes the hyperparameters for each machine learning algorithm, default values, and their ranges, respectively. We have chosen hyperparameters suitable for our study.

- For RF, we have taken n_estimators, max-depth, max_features, and criterion where n_estimators means the number of trees in the table, max_depth is the maximum number of trees, max_features are the maximum features to consider when searching for the best split, and criterion is the function that measures the quality of the split.
- In the case of AB, the hyperparameters chosen are: n_estimators are the maximum number of estimators where boosting is terminated, learning_rate is the weight applied to each machine learner at boosting iteration, and the algorithm is the real boosting algorithm to be used.
- For GNB, we use var_smoothing, the part of the largest variance of all features added to variances for calculation stability.
- In the case of SVM, we have used the "c" regularization parameter and kernel that describes the kernel type of the algorithm.
- For KNN, n_neighbours is the number of neighbors, weights are the weight function that describes the weights of neighbors, and leaf_size is passed to the algorithms like BallTree or KDTree.
- For SMOTE, Adasyn, and BL-SMOTE, k-neighbors are the number of neighbors, and sampling_strategy to resample the dataset are the hyperparameters to be used.

Table 5. Hyperparameter search space

Machine Learning Algorithms	Hyperparameters	Default	Range
Random Forest(RF)	n_estimators	100	[10–150]
	max_depth	None	[5–50]
	max_features	None	[0.01–1.0]
	criterion	”gini”	[“gini”, “entropy”]
AdaBoost(AB)	n_estimators	50	[50–500]
	learning_rate	1.0	[0.01–2.0]
	algorithm	“SAMME.R”	[“SAMME”, “SAMME.R”]
Gaussian Naive Bayes(GNB)	var_smoothing	1e-09	[0.0–1.0]
Support Vector Machine (SVM)	c	1.0	[1e-6, 100.0]
	kernel	“rbf”	[“linear”, “poly”, “rbf”, “sigmoid”]
k-Nearest Neighbour (KNN)	n_neighbours	5	[1–20]
	leaf_size	30	[10–100]
	weights	“uniform”	[“uniform”, “distance”]
SMOTE, ADASYN, BL-SMOTE	k_neighbours	5	[1–12]
	sampling_strategy	“not majority”	[“minority”, “not minority”, “not majority”, “all”]
Tomek Links(TL), Edited Nearest Neighbor(ENN)	sampling_strategy	“not majority”	[“majority”, “not minority”, “not majority”, “all”]
SMOTE + TL	sampling_strategy	“not majority”,	[“minority”, “not minority”, “not majority”, “all”]
	SMOTE	None,	[1, 12], [“minority”, “not minority”, “not majority”, “all”]
	TL	None	[“majority”, “not minority”, “not majority”, “all”]
SMOTE + ENN	sampling_strategy	“not majority”,	[“minority”, “not minority”, “not majority”, “all”]
	SMOTE	None,	[1, 12], [“minority”, “not minority”, “not majority”, “all”]
	ENN	None	[“majority”, “not minority”, “not majority”, “all”]

– In the case of SMOTE + ENN, the hyperparameters are the SMOTE object, ENN object, and sampling_strategy.
– Further, SMOTE + TL uses SMOTE object, Tomek links object, and sampling_strategy as hyperparameters.

HPO problem is defined as [61]:

$$x' = \arg\min f(x), \quad x \in \chi \tag{1}$$

where $f(x)$ is the objective function, χ is the hyperparameter search space, x is the set of best hyperparameters, and x can choose any hyperparameter from χ. HPO process is depicted in Figure 1. This paper has considered a single objective function to be minimized therefore only one performance metric is considered. To optimize more metrics multi-objective function is used.

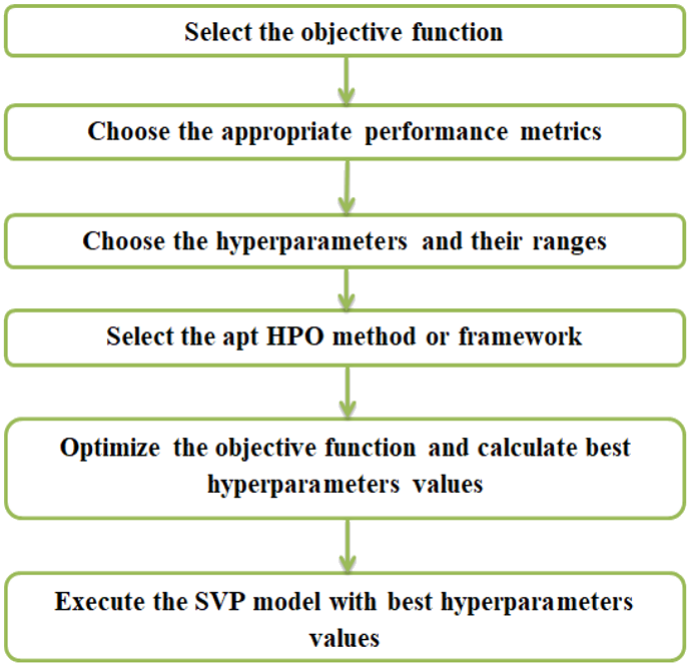

Figure 1. HPO process

3.5. Performance evaluation metrics

The accuracy metric fails in evaluating the SVP models in the imbalanced domain. Hence, we have used the area under the Receiver Operator Characteristic (ROC) curve (AUC) and $F1$-Score as the performance metric for the current study. ROC is a probability curve that plots the true positive rate ($TPrate$) against the false positive rate ($FPrate$). AUC gives the probability that the positive sample is ranked higher than the negative sample by the classifiers and is described in Figure 2. $TPrate$ is defined as the probability that the actual positive samples are correctly tested as positive whereas $FPrate$ is defined as the probability that the negative samples are tested as positive. AUC can be measured in terms of true positive rate ($TPrate$) and false positive rate ($FPrate$) as [62]:

$$Recall = TPrate = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{2}$$

$$Precision = \frac{TP}{TP + FN} \tag{3}$$

$$FPrate = \frac{FP}{FP + TN} = \frac{FP}{N} \tag{4}$$

$$AUC = \frac{1 + TPrate - FPrate}{2} \tag{5}$$

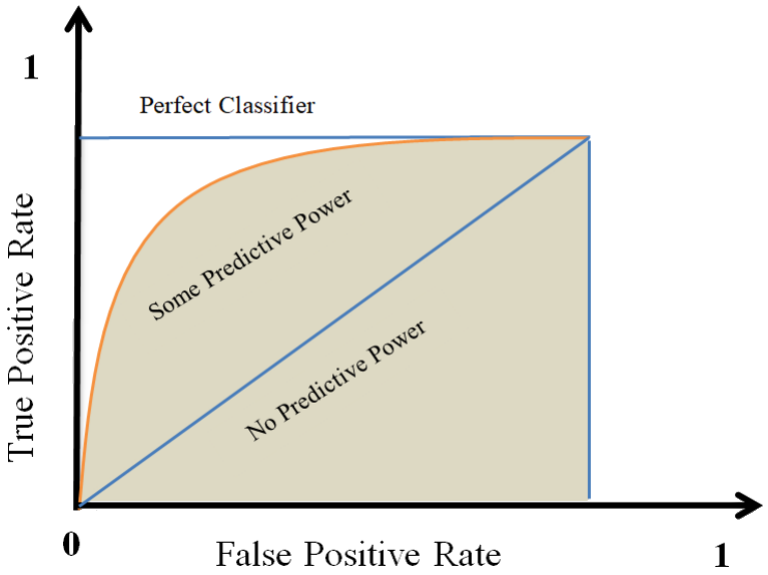

Figure 2. Area under ROC curve

The performance values of AUC range from 0 to 1 and these values figure out how well are classifiers at distinguishing between positive and negative classes. The high-performance model has AUC close to 1, whereas the low-performance model has AUC close to 0.5 [63]. $F1$-Score is defined as the harmonic mean of precision and recall each weighted equally.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

3.6. Data complexity measures

It is observed that the efficiency of the machine learning algorithms is not only degraded by imbalanced datasets but also by the degree of class overlapping. The degree of class overlapping is calculated by the data complexity measures [36] and [37].

The current study has focused on the feature-based measure (maximum Fisher's discriminant ratio ($F1$) and class imbalance measure (imbalance ratio). $F1$ measures the maximum discriminative power of all features in different classes and is computed as [38]:

$$F1 = \frac{(\mu c1 - \mu c2)^2}{\sigma c1^2 + \sigma c2^2} \tag{7}$$

where μci and σci are the mean and variance of values of the class i feature, respectively. The higher values of $F1$ indicate lower complexity; henceforth the classification problem is simple. The lower value of $F1$ shows that classes are highly overlapped. The imbalance ratio (IR) in the case of binary classification is calculated as [64]:

$$IR = \frac{No.\ of\ majority\ class\ instances}{No.\ of\ minority\ class\ instances} \tag{8}$$

4. Experimental framework

This section includes an Experimental procedure (Section 4.1) and a statistical test (Section 4.2). Figure 3 shows the experimental framework that illustrates the working of the model [65].

4.1. Experimental procedure

The experimental procedure (see Table 6) is based on the pseudo-code presented in [3]. The experimental methodology is illustrated through Algorithm1 and Subalgorithm.

Table 6. Summary of the information regarding the experimental setup of the current study

Datasets	Machine Learning Techniques	Resampling Techniques	Hyperparameter Optimization Method	Performance Metrics	Data Complexity Measures
Drupal, Moodle, PHP-MyAdmin	RF, AB, GNB, KNN, SVM	SMOTE, ADASYN, BL SMOTE, SMOTE + TL, SMOTE + ENN	Optuna	AUC and $F1$-Score	$F1$ and IR

– Algorithm1 calculates the AUC value and $F1$-Score of each scenario for HPO explained in Table 1. The experiments are performed on three datasets given in Table 4.
– The methodology uses three-fold cross-validation repeated 50 times to maintain the random order construed in Subalgorithm. The three-fold cross-validation splits the dataset into three parts (2:1) where two parts are used for training and one part is used as the testing dataset. In other words, the training dataset is 66.66% of the entire dataset and the testing dataset is 33.33% of the whole dataset. In Algorithm1, lines 5–8 calculates evaluation metrics for (Ad + Rn) and (Ao + Rn) scenarios and lines 9–12 calculates evaluation metrics for (Ad + Rd), (Ao + Rd), (Ad + Ro), and (Ao + Ro) scenarios.
– There are two arrays default[] and optimized[] which stores the hyperparameters of machine learners and resamplers. Default[] indicates that the machine learners will run in their default settings whereas the optimized[] array is calculated by performing hyperparameter tuning mentioned in Section 3.4.
– For hyperparameter tuning, there is a need for a validation dataset. The training dataset (66.66%) is further split into two parts, i.e., new training dataset and one part of the validation dataset which means 44.435% is the new training dataset and 22.217% is the validation dataset. The training dataset trains the classifier with chosen

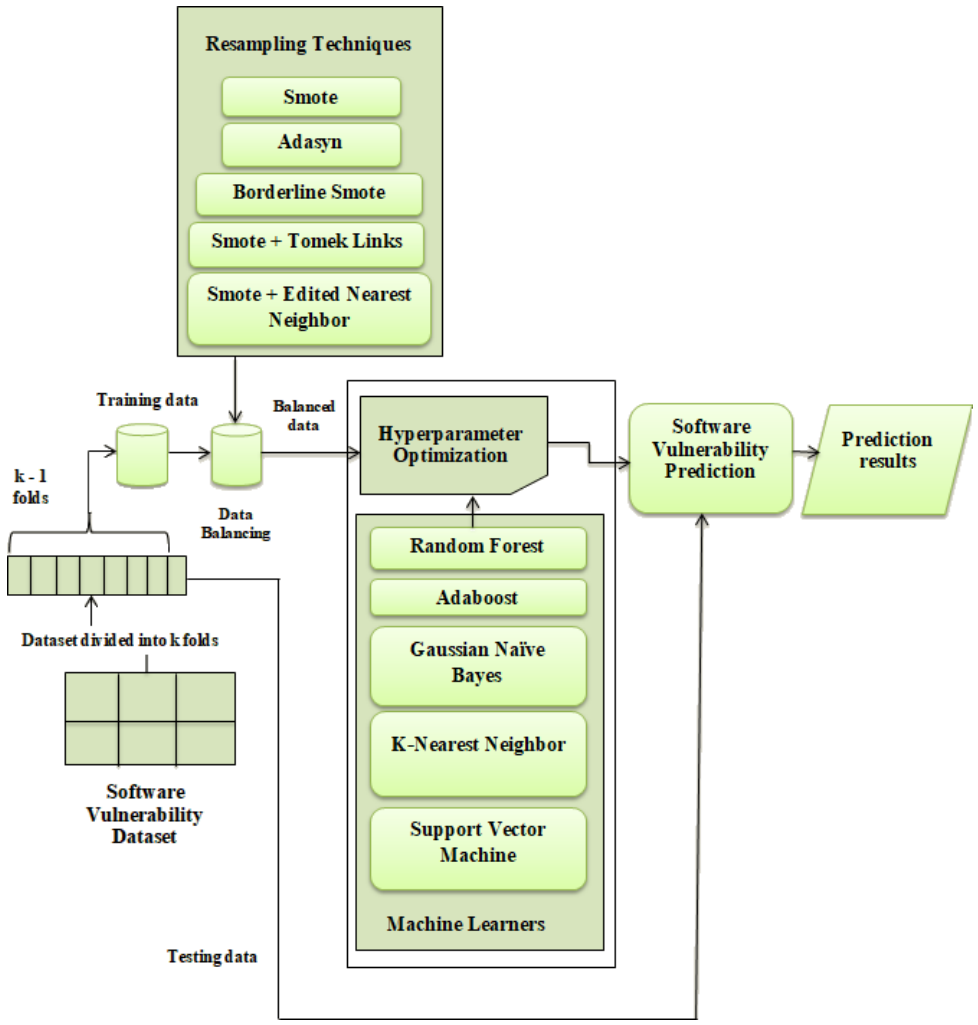

Figure 3. Working of proposed methodology

```
In [3]: study = optuna.create_study(direction='maximize')
        study.optimize(objective, n_trials=100)

[I 2022-09-12 20:45:51,610] Trial 91 finished with value: 0.840579208548848 and parameters: {'n_estimators': 133, 'max_dept
h': 41, 'max_features': 0.9009689339823158, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:52,312] Trial 92 finished with value: 0.8415460651513782 and parameters: {'n_estimators': 129, 'max_dept
h': 47, 'max_features': 0.8534884394041747, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:53,172] Trial 93 finished with value: 0.8435169623404918 and parameters: {'n_estimators': 141, 'max_dept
h': 46, 'max_features': 0.978168098981247, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:53,898] Trial 94 finished with value: 0.8423431933868365 and parameters: {'n_estimators': 122, 'max_dept
h': 45, 'max_features': 0.945286260751885, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:54,317] Trial 95 finished with value: 0.8383656844947168 and parameters: {'n_estimators': 70, 'max_dept
h': 37, 'max_features': 0.8398785902121185, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:55,094] Trial 96 finished with value: 0.8453117056532616 and parameters: {'n_estimators': 147, 'max_dept
h': 25, 'max_features': 0.5457116291238912, 'criterion': 'gini'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:55,736] Trial 97 finished with value: 0.8226634739917472 and parameters: {'n_estimators': 133, 'max_dept
h': 6, 'max_features': 0.4952815803481252, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.

[I 2022-09-12 20:45:56,074] Trial 98 finished with value: 0.8344033357885351 and parameters: {'n_estimators': 62, 'max_dept
h': 10, 'max_features': 0.371176569734877, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.
[I 2022-09-12 20:45:56,719] Trial 99 finished with value: 0.8513266121046007 and parameters: {'n_estimators': 126, 'max_dept
h': 42, 'max_features': 0.42050213915806467, 'criterion': 'entropy'}. Best is trial 88 with value: 0.8593560434926659.

In [4]: trial = study.best_trial
        print('AUC: {}'.format(trial.value))

        AUC: 0.8593560434926659

In [5]: print("Best hyperparameters: {}".format(trial.params))

        Best hyperparameters: {'n_estimators': 133, 'max_depth': 43, 'max_features': 0.8915329694493298, 'criterion': 'entropy'}
```

Figure 4. Example of the result of RF hyperparameter tuning on Drupal dataset

Algorithm 1: Experimental Methodology

Input: D: A set of software vulnerability datasets, D = {PHPMyAdmin, Moodle, and Drupal}
R: A set of resampling techniques, R = {unbalanced, Smote, Adasyn, BLSmote, SmoteTL, SmoteENN}
L: A set of machine learning methods, L = {RF, AB, GNB, KNN, SVM}
RC:A set of resampling conditions, RC = {none, default, optimized}
HP:A set of arrays containing default hyperparameters and optimized hyperparameters, HP = {default [], optimized []}
Output: AUC and F1-Score performance of a combination of learners and resamplers
Begin:

```
1. for each data in D do
2.     for each l in L do
3.         for each r in R do
4.             for each rc in RC do
5.                 if(balancer == unbalanced && rc == none)
6.                     M1= Subalgorithm (HP = default [], l, data, r, rc)
7.                     M2= Subalgorithm (HP = optimized [], l, data, r, rc)
8.                 end if
9.                 if(balancer!=unbalanced && ((rc == default || rc == optimized))
10.                    M1= Subalgorithm (HP = default [], l, data, r, rc)
11.                    M2= Subalgorithm (HP = optimized [], l, data, r, rc)
12.                end if
13.            end for
14.        end for
15.    end for
16. end for
```

End

Subalgorithm: learners and resamplers performing with default or optimized hyperparameters

Input: HP = default [] or optimized [] , l: l ϵ L, data: data ϵ D, r: r ϵ R, rc: rc ϵ RC
M = 50, N = 3 // N fold cross-validation M times
Output: mean AUC metric value and F1-Score
Begin:

```
1. repeat (M times)
2.     randomized order from data
3.     generate k bins from data
4.     for each k in N do
5.         testData = bin (k)
6.         trainingData = data – testData
7.         new_data = r (trainingData)                    //data resampling
8.         predictor = l (new_data)
9.         apply predictor to testData
10.        return mean AUC metric value and F1-Score
11.    end for
12. end repeat
```

End

hyperparameters and their ranges. An objective function is optimized (AUC in our case), and on the validation part, 100 iterations are performed to achieve the best hyperparameters. Figure 4 describes the example of hyperparameter tuning of RF on the Drupal dataset for scenario 2.

- Further, the best hyperparameters are evaluated on the testing dataset using subalgorithm which calculates the AUC and $F1$-Score of each scenario. The final ratio of the training, validation, and testing part is 1.33:0.66:1.
- This paper focuses on optimizing a single objective function which means we are focusing on maximizing the AUC performance metric using the best hyperparameters configuration.
- Subalgorithm takes the dataset, hyperparameters (default or optimized), machine learning algorithm, resampling technique, and resampling condition as input and returns the AUC metric and $F1$-score as output to Algorithm 1.

4.2. Statistical tests

The current study has applied Wilcoxon signed-rank test [66, 67] for statistical analysis. The Wilcoxon signed-rank test is a non-parametric significance test, usually applied when the

readings are not normally distributed. It is used to test the differences in the performance (AUC) of six scenarios mentioned in Table 1. The p-value indicates a significant difference in the readings. Since six statistical tests are being performed so the p-value will now be changed to $0.05/6 = 0.0083$ as per the Bonferroni correction method [68, 69]. To conduct this test, the SPSS tool is used. The null hypothesis (Ho) describes that the performance values are equal and the alternate hypothesis (Ha) indicates that performance value differs.

To answer the research questions comparisons of the following cases are required:

1. (AdRn) & (AoRn): Default learner parameters with no resampling and optimized hyperparameters of learner with no resampling.
2. (AdRn) & (AdRd): Default learner parameters with no resampling and Default learner parameters with default resampler parameters.
3. (AdRd) & (AoRd): Default learner parameters with default resampling and optimized hyperparameters of learner with default resampler parameters.
4. (AdRd) & (AdRo): Default learner parameters with default resampling and default learner parameters with optimized hyperparameters of resampler.
5. (AdRn) & (AoRo): Default learner parameters with no resampling and optimized hyperparameters of both learner and resampler.
6. (AdRd) & (AoRo): Default learner parameters with default resampler parameters and optimized hyperparameters of both learner and resampler.

5. Results and analysis

This section explains the experimental results for the six scenarios of HPO. Further, the analysis of the results is performed to check whether each scenario is statistically significant. The results are based on the experimental procedure mentioned in Section 4. The goal of this paper is not to find the best resampler or best machine learner but to find the impact of dual HPO on SVP models.

5.1. AUC results

Tables 7–11 present the AUC values for each HPO scenario explained in Table 1. The blue-shaded cells indicate the highest AUC value per row. The yellow-shaded cell indicates the highest $F1$-Score among all scenarios. The bold + shaded cell indicates the highest AUC per dataset for each machine learner. It should be noted that our study works on the single objective function which has optimized the AUC metric and we are measuring AUC improvements. Although we have shown $F1$-Score in the results HPO and dual HPO will affect this metric mainly when used in the objective function which is a multi-objective problem that is out of the scope of this paper. The paper has still presented the effect of HPO and dual HPO on $F1$-Score when the single-objective function is optimized.

In Table 7, it has been observed that:

– In the case of RF, for Drupal, scenario 6 has achieved the highest AUC value of 0.8461 with SMOTE resampling technique.
– For Moodle dataset, BL SMOTE for scenario 6 has shown the highest AUC of 0.7783.
– For PHPMyAdmin, scenario 3 with default resampling results in a maximum AUC value of 0.7081.

- In the case of Drupal $F1$-Score has the highest value of 0.7088 in scenario 6 of Adasyn. For Moodle, $F1$-Score has the highest value of 0.0664 in scenario 6 of SMOTE. In PHPMyAdmin, the highest value of 0.2594 is attained in scenario 1.

In Table 8,

- For AB, it is observed that BL SMOTE for scenario 6 has performed the best for Drupal, with an AUC value of 0.8524.
- For Moodle, Adasyn for scenario 6 has resulted best, with an AUC value of 0.8402.
- In the case of PHPMyAdmin, scenario 5 has performed the best with an AUC value of 0.7351.
- $F1$-Score in the case of Drupal is highest with a value of 0.7566 for BL SMOTE scenario 6. Moodle has the highest $F1$-Score with a value of 0.0778 in SMOTE + TL scenario 6. PHPMyAdmin has the highest value of $F1$-Score 0.2881 in scenario 4 of SMOTE + TL.

Table 9 gives the AUC results for the GNB algorithm,

- SMOTE + TL with scenario 6 has provided the highest AUC value of 0.8777 for the Drupal dataset.
- For Moodle, SMOTE + ENN in scenario 6 has given the highest AUC value of 0.8909.
- In the case of PHPMyAdmin, Adasyn in scenario 4 has performed the best with the maximum AUC value of 0.7561.
- $F1$-Score in the case of Drupal is highest with a value of 0.5777 for SMOTE + ENN scenario 5. Moodle has the highest $F1$-Score with a value of 0.1679 in SMOTE scenario 6. PHPMyAdmin has the highest value of $F1$-Score 0.2648 in scenario 1.

Table 10 presents the AUC values for the KNN algorithm,

- For the Drupal dataset, SMOTE in scenario 6 has performed the best, with an AUC value of 0.8652.
- In the case of Moodle, BL SMOTE in scenario 6 has performed the best, with an AUC value of 0.7997.
- For PHPMyAdmin, the highest AUC value of 0.7001 is achieved by SMOTE in scenario 4.
- $F1$-Score in the case of Drupal is highest with a value of 0.6571 for SMOTE + ENN scenario 5. Moodle has the highest $F1$-Score with a value of 0.0743 in Adasyn scenario 5. PHPMyAdmin has the highest value of $F1$-Score 0.2683 in BL SMOTE scenario 6.

Table 11 gives the AUC performance value measure for the SVM algorithm,

- For the Drupal dataset, SMOTE + TL in scenario 4 has performed the best, with an AUC value of 0.8825.
- In the case of Moodle, BL SMOTE in scenario 6 has the best AUC value of 0.8492.
- For PHPMyAdmin, SMOTE + TL in scenario 4 has given the best AUC value of 0.7101.
- $F1$-Score in the case of Drupal is highest with a value of 0.6155 for BL SMOTE scenario 5. Moodle has the highest $F1$-Score with a value of 0.1133 in SMOTE scenario 4. PHPMyAdmin has the highest value of $F1$-Score 0.3872 in scenario 4.

It has been found that there lies a slight difference among the results of resampling techniques per HPO scenario; therefore we have not compared them in the study. Figures 5–7 describe the average AUC performance values of each HPO scenario calculated column-wise for Drupal, Moodle, and PHPMyAdmin, respectively.

Table 7. Performance values for RF algorithm

Resampling techniques	Dataset	Scenarios											
		$A_d + R_n$		$A_o + R_n$		$A_d + R_d$		$A_o + R_d$		$A_d + R_o$		$A_o + R_o$	
		AUC	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score
NONE	PHP MyAdmin	0.6879	0.2594	0.6274	0.2327								
SMOTE						0.7025	0.2308	0.7044	0.1054	0.6741	0.09	0.6460	0.1269
ADASYN						0.7005	0.2389	0.6568	0.1658	0.6964	0.1421	0.6286	0.2208
BL SMOTE						0.7036	0.2323	0.6433	0.1299	0.6084	0.0791	0.6159	0.1365
SMOTE + TL						**0.7081**	0.2329	0.6302	0.1228	0.7029	0.0282	0.6677	0.0083
SMOTE + ENN						0.6515	0.2068	0.6965	0.0730	0.6164	0.0197	0.6863	0.1196
NONE	Moodle	0.6341	0.0	0.5644	0.0								
SMOTE						0.7264	0.0289	0.7057	0.0594	0.7169	0.0181	0.7374	0.0664
ADASYN						0.7358	0.0295	0.7019	0.0310	0.7042	0.0188	0.7559	0.0414
BL SMOTE						0.6641	0.0191	0.7033	0.0302	0.7149	0.0012	0.7316	0.0275
SMOTE + TL						0.7331	0.0286	0.7252	0.0032	0.7043	0.0057	**0.7783**	0.0244
SMOTE + ENN						0.7401	0.0281	0.6709	0.0222	0.7159	0.0574	0.7622	0.0263
NONE	Drupal	0.8131	0.5684	0.8145	0.4108								
SMOTE						0.8024	0.6087	0.7966	0.5527	0.8302	0.6458	0.8416	0.5742
ADASYN						0.7994	0.6082	0.8011	0.5427	0.8021	0.6141	0.8190	0.7088
BL SMOTE						0.7967	0.6022	0.8221	0.6388	0.8037	0.6328	0.8353	0.6085
SMOTE + TL						0.8019	0.6166	0.8157	0.6032	0.8277	0.5761	0.8236	0.5239
SMOTE + ENN						0.7961	0.6093	0.8196	0.5568	0.8131	0.5537	**0.8461**	0.4776

Table 8. Performance values for AB algorithm

Resampling techniques	Dataset	Scenarios											
		$A_d + R_n$		$A_o + R_n$		$A_d + R_d$		$A_o + R_d$		$A_d + R_o$		$A_o + R_o$	
		AUC	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score
NONE	PHPMyAdmin	0.6357	0.2799	0.6331	0.0484								
SMOTE						0.6095	0.1863	0.6895	0.1346	0.7189	0.1785	0.6882	0.0743
ADASYN						0.6136	0.1777	0.6825	0.2451	0.6769	0.2072	0.6206	0.1968
BL SMOTE						0.6113	0.1974	0.6311	0.241	0.7121	0.3295	0.6282	0.0287
SMOTE + TL						0.6107	0.1844	0.6584	0.2881	0.6523	0.1226	0.6145	0.0427
SMOTE + ENN						0.6081	0.1829	0.6621	0.2751	**0.7351**	0.1922	0.6419	0.1829
NONE	Moodle	0.7307	0.0009	0.8346	0.0								
SMOTE						0.7403	0.0381	0.7773	0.0261	0.8153	0.0391	0.7941	0.0434
ADASYN						0.7424	0.2469	0.8191	0.0391	0.7721	0.0353	**0.8402**	0.0402
BL SMOTE						0.7199	0.0514	0.8085	0.0209	0.8022	0.0318	0.8177	0.0407
SMOTE + TL						0.7434	0.0358	0.8239	0.0610	0.7746	0.0385	0.7611	0.0778
SMOTE + ENN						0.7521	0.0411	0.7829	0.0118	0.7483	0.0287	0.7737	0.0838
NONE	Drupal	0.7729	0.5177	0.8082	0.6460								
SMOTE						0.7655	0.5561	0.8355	0.5259	0.8065	0.6156	0.8175	0.6768
ADASYN						0.7643	0.5573	0.8239	0.6303	0.7609	0.4338	0.8511	0.7042
BL SMOTE						0.7588	0.5575	0.8083	0.5934	0.7962	0.5428	**0.8524**	0.7566
SMOTE + TL						0.7718	0.5739	0.7909	0.5875	0.7995	0.5291	0.8501	0.6203
SMOTE + ENN						0.7515	0.6043	0.7631	0.6561	0.7432	0.4527	0.8299	0.6324

Table 9. Performance values for GNB algorithm

Resampling techniques	Dataset	Scenarios											
		$A_d + R_n$		$A_o + R_n$		$A_d + R_d$		$A_o + R_d$		$A_d + R_o$		$A_o + R_o$	
		AUC	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score
NONE	PHPMyAdmin	0.6941	0.2648	0.7029	0.2206								
SMOTE						0.6851	0.2482	0.7265	0.0779	0.6221	0.2392	0.6975	0.1581
ADASYN						0.6842	0.2442	**0.7561**	0.2265	0.6582	0.2133	0.6005	0.1152
BL SMOTE						0.6525	0.2421	0.7007	0.0759	0.6268	0.2397	0.6671	0.0927
SMOTE + TL						0.6847	0.2547	0.7538	0.0762	0.6873	0.2869	0.6752	0.1522
SMOTE + ENN						0.7006	0.2452	0.7345	0.0545	0.6571	0.0605	0.6043	0.2357
NONE	Moodle	0.8151	0.0782	0.8453	0.0128								
SMOTE						0.8075	0.0606	0.8310	0.0721	0.7864	0.0291	0.8775	0.1679
ADASYN						0.8039	0.0602	0.8412	0.0356	0.7329	0.0182	0.8833	0.1114
BL SMOTE						0.8024	0.0883	0.8167	0.0410	0.7748	0.1454	0.8691	0.1271
SMOTE + TL						0.8094	0.0629	0.8313	0.0756	0.7668	0.0622	0.8553	0.0721
SMOTE + ENN						0.8040	0.0593	0.8404	0.0274	0.7962	0.0273	**0.8909**	0.1261
NONE	Drupal	0.7756	0.4924	0.8322	0.2132								
SMOTE						0.7781	0.4957	0.8456	0.5164	0.7926	0.5247	0.8735	0.4383
ADASYN						0.7723	0.5060	0.8596	0.4924	0.7533	0.4617	0.8482	0.5325
BL SMOTE						0.7731	0.5121	0.8236	0.5616	0.8226	0.5777	0.8551	0.4299
SMOTE + TL						0.7795	0.5034	0.8194	0.4556	0.7594	0.5009	**0.8777**	0.5396
SMOTE + ENN						0.7591	0.5626	0.8198	0.5231	0.7931	0.6211	0.8721	0.5521

Table 10. Performance values for KNN algorithm

Resampling techniques	Dataset	Scenarios											
		$A_d + R_n$		$A_o + R_n$		$A_d + R_d$		$A_o + R_d$		$A_d + R_o$		$A_o + R_o$	
		AUC	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score
NONE	PHPMyAdmin	0.6626	0.2099	0.6274	0.2327								
SMOTE						0.6676	0.2010	0.6746	0.157	0.6413	0.1956	0.6367	0.074
ADASYN						0.6504	0.2044	**0.7001**	0.1866	0.6194	0.1763	0.6194	0.2077
BL SMOTE						0.6508	0.1931	0.6674	0.084	0.6596	0.2455	0.6784	0.2683
SMOTE + TL						0.6598	0.2046	0.6657	0.2097	0.6764	0.2616	0.6661	0.1829
SMOTE + ENN						0.6676	0.1946	0.6561	0.2029	0.6561	0.1772	0.6432	0.1993
NONE	Moodle	0.5199	0.0	0.5644	0.0								
SMOTE						0.6917	0.0411	0.5683	0.023	0.7219	0.0372	0.7371	0.0566
ADASYN						0.6972	0.0389	0.6445	0.0025	0.7441	0.0743	0.7364	0.0431
BL SMOTE						0.6112	0.0689	0.5842	0.002	0.7133	0.0267	**0.7997**	0.0726
SMOTE + TL						0.7058	0.0411	0.5862	0.037	0.5328	0.0247	0.7458	0.043
SMOTE + ENN						0.6901	0.0382	0.5972	0.0132	0.7306	0.0646	0.7206	0.0563
NONE	Drupal	0.7679	0.5193	0.8145	0.4108								
SMOTE						0.7619	0.5938	0.7276	0.5709	0.8128	0.4338	0.8351	0.6154
ADASYN						0.7596	0.6035	0.7984	0.6258	0.8067	0.6055	0.8652	0.4401
BL SMOTE						0.7512	0.6034	0.7742	0.5148	0.7561	0.4055	0.8624	0.6039
SMOTE + TL						0.7582	0.5958	0.8053	0.5483	0.7729	0.7082	0.8124	0.5026
SMOTE + ENN						0.7501	0.6151	0.8055	0.6413	0.8143	0.6751	0.7626	0.5665

Table 11. Performance values for SVM algorithm

Resampling techniques	Dataset	Scenarios											
		$A_d + R_n$		$A_o + R_n$		$A_d + R_d$		$A_o + R_d$		$A_d + R_o$		$A_o + R_o$	
		AUC	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score	*AUC*	*F*1-Score
NONE	PHPMyAdmin	0.6354	0.2999	0.6629	0.2215								
SMOTE						0.6973	0.2444	0.6536	0.2738	0.6251	0.1862	0.6284	0.0
ADASYN						0.6905	0.2311	0.6572	0.2120	0.6951	0.0	0.6068	0.2292
BL SMOTE						0.6881	0.2465	0.6535	0.3872	0.6321	0.0	0.6108	0.0019
SMOTE + TL						0.6934	0.2474	**0.7101**	0.3218	0.6113	0.0	0.6625	0.1988
SMOTE + ENN						0.6267	0.1836	0.6311	0.1048	0.6923	0.3553	0.6049	0.1678
NONE	Moodle	0.4028	0.0	0.7506	0.0								
SMOTE						0.7303	0.0386	0.7923	0.1133	0.8415	0.0224	0.8236	0.1129
ADASYN						0.7379	0.0379	0.8265	0.0161	0.7885	0.0223	0.8202	0.0099
BL SMOTE						0.7721	0.0622	0.8215	0.0146	0.7628	0.0512	**0.8492**	0.0598
SMOTE + TL						0.7536	0.0379	0.8210	0.0538	0.8086	0.0176	0.7810	0.0311
SMOTE + ENN						0.7557	0.0365	0.7765	0.0091	0.7505	0.0652	0.8392	0.0801
NONE	Drupal	0.6955	0.4423	0.8001	0.1112								
SMOTE						0.8055	0.5681	0.8107	0.5831	0.8218	0.5721	0.8126	0.3564
ADASYN						0.7942	0.5805	0.8369	0.2854	0.8141	0.5803	0.8704	0.4923
BL SMOTE						0.7732	0.5934	0.8396	0.1655	0.8217	0.6155	0.8553	0.2401
SMOTE + TL						0.8001	0.5611	0.7744	0.4553	0.7976	0.4847	**0.8825**	0.3454
SMOTE + ENN						0.7931	0.5592	0.7859	0.5776	0.8052	0.5432	0.8518	0.4882

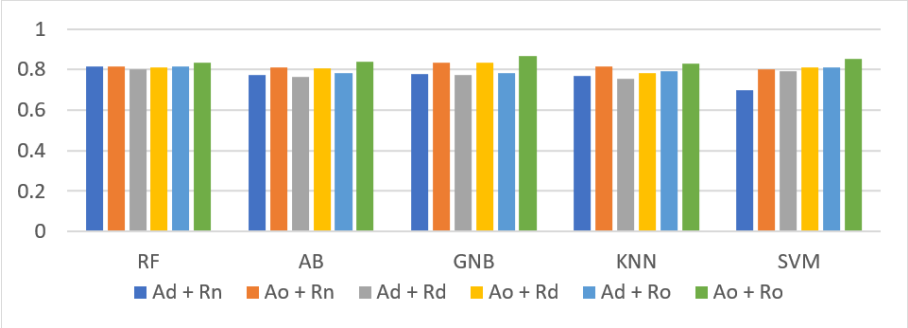

Figure 5. Average *AUC* performance values of each HPO scenario for the Drupal dataset

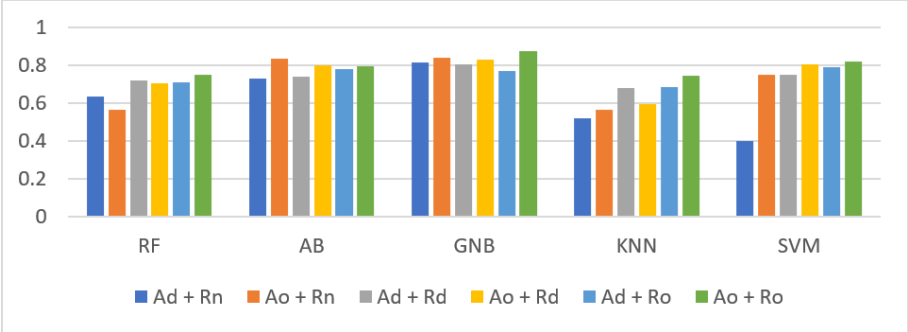

Figure 6. Average *AUC* performance values of each HPO scenario for the Moodle dataset

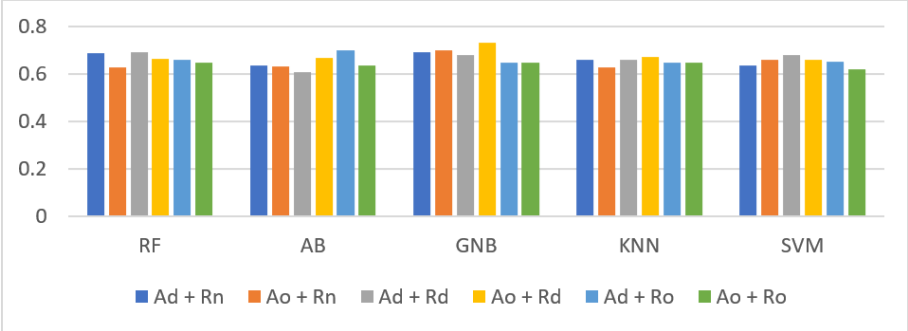

Figure 7. Average *AUC* performance values of each HPO scenario for the PHPMyAdmin dataset

5.2. Statistical results

Tables 12–14 show the results of the Wilcoxon signed-rank test applied to six scenarios for Drupal, Moodle, and PHPMyAdmin, respectively. The grey shaded area shows significant improvement, and "*" indicates the *AUC* value has decreased.

Table 12 shows the *p*-values, calculated after comparing the six cases mentioned above, for the Drupal dataset. The following observations depict the significant improvements of each machine learning algorithm.

– RF shows improvement in case 4 SMOTE, case 5 SMOTE, BL SMOTE, and SMOTE + ENN, and in case 6.

Table 12. Statistical comparison results among all five machine learners for Drupal dataset

Scenarios	Machine learning algorithm	Resampling techniques					
		NONE	SMOTE	ADASYN	BL SMOTE	SMOTE + TL	SMOTE + ENN
(A_dR_n) & (A_oR_n)	RF	0.211					
(A_dR_n) & (A_dR_d)			*	*	*	*	*
(A_dR_n) & (A_oR_d)			*	*	0.059	0.331	0.173
(A_dR_n) & (A_dR_o)			0.066	*	*	0.022	0.873
(A_dR_n) & (A_oR_o)			<0.001	0.511	0.002	0.039	<0.001
(A_dR_d) & (A_oR_o)			<0.001	0.005	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_n)	AB	<0.001					
(A_dR_n) & (A_dR_d)			*	*	*	*	*
(A_dR_n) & (A_oR_d)			<0.001	<0.001	<0.001	0.057	*
(A_dR_n) & (A_dR_o)			<0.001	*	0.004	<0.001	*
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_n)	GNB	<0.001					
(A_dR_n) & (A_dR_d)			0.930	*	*	0.360	*
(A_dR_n) & (A_oR_d)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_dR_o)			0.023	*	<0.001	*	0.018
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_n)	KNN	<0.001					
(A_dR_n) & (A_dR_d)			*	*	*	*	*
(A_dR_n) & (A_oR_d)			<0.001	0.002	0.199	<0.001	<0.001
(A_dR_n) & (A_dR_o)			0.023	<0.001	*	0.705	<0.001
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	0.520
(A_dR_d) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.036
(A_dR_n) & (A_oR_n)	SVM	<0.001					
(A_dR_n) & (A_dR_d)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_d)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_dR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			0.037	<0.001	<0.001	<0.001	<0.001

- For AB, case 1, case 3 SMOTE, Adasyn, and BL SMOTE, case 4 SMOTE, BL SMOTE, SMOTE + TL, case 5, and 6 shows improvement.
- GNB shows improvement for scenarios 1, 3, 5, and 6 and BL SMOTE in case 4.
- KNN has complete improvements for case 1, and case 3 except for BL SMOTE, case 4 Adasyn and SMOTE + ENN, case 5 and 6 except SMOTE + ENN.
- SVM shows improvement in all the cases except for case 6 of SMOTE.

Table 13 shows the p-values for the Moodle dataset. The observations are as follows:

- RF has shown improvement in cases 2, 4, 5, case 3 except SMOTE + ENN, and case 6 with BL SMOTE and SMOTE + TL.
- AB shows improvement for case 1, case 5, case 2 SMOTE + ENN, case 3 except SMOTE + ENN, case 4 except SMOTE + ENN, and case 6 except SMOTE + TL and SMOTE + ENN.
- GNB shows improvement for case 1, case 5, case 6, for case 3 SMOTE + ENN.
- KNN shows improvement in case 1, case 2, case 5, case 3 except Adasyn and SMOTE + ENN, case 4 except SMOTE + TL, and case 6 except Adasyn and SMOTE + ENN.
- SVM shows improvement in all cases except SMOTE + ENN of the case 6.

Table 13. Statistical comparison results among all five machine learners for Moodle dataset

Scenarios	Machine learning algorithm	Resampling techniques					
		NONE	SMOTE	ADASYN	BL SMOTE	SMOTE + TL	SMOTE + ENN
(A_dR_n) & (A_oR_n)	RF	*					
(A_dR_n) & (A_dR_d)			<0.001	<0.001	0.003	<0.001	<0.001
(A_dR_n) & (A_oR_d)			<0.001	<0.001	<0.001	0.002	0.047
(A_dR_n) & (A_dR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			0.071	0.036	<0.001	0.004	0.709
(A_dR_n) & (A_oR_n)	AB	<0.001					
(A_dR_n) & (A_dR_d)			0.183	0.172	*	0.159	0.004
(A_dR_n) & (A_oR_d)			0.007	<0.001	<0.001	<0.001	0.028
(A_dR_n) & (A_dR_o)			0.003	0.001	<0.001	0.002	0.023
(A_dR_n) & (A_oR_o)			<0.001	<0.001	0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			<0.001	<0.001	<0.001	0.025	0.037
(A_dR_n) & (A_oR_n)	GNB	<0.001					
(A_dR_n) & (A_dR_d)			*	*	*	*	*
(A_dR_n) & (A_oR_d)			0.015	0.229	0.015	0.150	<0.001
(A_dR_n) & (A_dR_o)			*	*	*	*	*
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	0.004	<0.001
(A_dR_d) & (A_oR_o)			<0.001	<0.001	<0.001	0.002	<0.001
(A_dR_n) & (A_oR_n)	KNN	<0.001					
(A_dR_n) & (A_dR_d)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_d)			<0.001	0.031	<0.001	<0.001	0.0153
(A_dR_n) & (A_dR_o)			0.004	<0.001	<0.001	0.785	<0.001
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			<0.001	0.022	<0.001	0.004	<0.013
(A_dR_n) & (A_oR_n)	SVM	<0.001					
(A_dR_n) & (A_dR_d)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_d)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_dR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_n) & (A_oR_o)			<0.001	<0.001	<0.001	<0.001	<0.001
(A_dR_d) & (A_oR_o)			0.022	<0.001	0.984	0.069	<0.001

Table 14 presents the p-values for the PHPMyAdmin dataset. The statistical results are:

- RF has shown no significant improvement.
- AB has shown significant improvement for case 3 Adasyn, case 4 except SMOTE + TL, SMOTE for case 5 and case 6.
- GNB has shown significant improvement for case 3 Adasyn
- KNN has shown significant improvement for case 1 and case 2 SMOTE + ENN.
- SVM shows a significant improvement in case 2 except SMOTE + ENN, case 3 SMOTE + TL, and case 4 Adasyn.

5.3. Illustration of research questions

Results reported in Tables 7–14 contribute to the answers to research questions.
RQ 1: How much is dual HPO effective in improving the performance of SVP models? The statistical comparisons of AdRn & AoRo (case 5 of Section 4.2) show the effectiveness of dual HPO on SVP models. There exist a total of 75 instances out of which 48 instances

Table 14. Statistical comparison results among all five machine learners for PHPMyAdmin dataset

Scenarios	Machine learning algorithm	Resampling techniques					
		NONE	SMOTE	ADASYN	BL SMOTE	SMOTE + TL	SMOTE + ENN
(A_dR_n) & (A_oR_n)	RF	*					
(A_dR_n) & (A_dR_d)			0.100	0.159	0.089	0.041	*
(A_dR_n) & (A_oR_d)			0.638	*	*	*	0.985
(A_dR_n) & (A_dR_o)			*	0.558	*	0.107	*
(A_dR_n) & (A_oR_o)			*	*	*	*	*
(A_dR_d) & (A_oR_o)			*	*	*	*	0.009
(A_dR_n) & (A_oR_n)	AB	*					
(A_dR_n) & (A_dR_d)			*	*	*	*	*
(A_dR_n) & (A_oR_d)			0.020	0.005	*	0.181	0.033
(A_dR_n) & (A_dR_o)			0.004	0.006	<0.001	0.517	<0.001
(A_dR_n) & (A_oR_o)			<0.001	*	*	0.237	0.742
(A_dR_d) & (A_oR_o)			<0.001	0.812	0.325	0.683	0.063
(A_dR_n) & (A_oR_n)	GNB	0.361					
(A_dR_n) & (A_dR_d)			*	*	*	*	0.541
(A_dR_n) & (A_oR_d)			0.059	<0.001	0.361	0.203	0.031
(A_dR_n) & (A_dR_o)			*	*	*	*	*
(A_dR_n) & (A_oR_o)			*	*	*	*	*
(A_dR_d) & (A_oR_o)			0.042	*	0.095	*	*
(A_dR_n) & (A_oR_n)	KNN	0.007					
(A_dR_n) & (A_dR_d)			0.512	*	*	*	<0.001
(A_dR_n) & (A_oR_d)			0.381	0.041	0.713	0.835	*
(A_dR_n) & (A_dR_o)			*	*	*	0.406	*
(A_dR_n) & (A_oR_o)			*	*	0.173	0.492	*
(A_dR_d) & (A_oR_o)			*	*	0.056	0.443	*
(A_dR_n) & (A_oR_n)	SVM	0.031					
(A_dR_n) & (A_dR_d)			<0.001	<0.001	<0.001	<0.001	0.567
(A_dR_n) & (A_oR_d)			0.195	0.080	0.207	<0.001	*
(A_dR_n) & (A_dR_o)			0.661	0.005	0.695	0.447	*
(A_dR_n) & (A_oR_o)			*	0.354	*	*	0.422
(A_dR_d) & (A_oR_o)			*	*	*	0.272	*

significantly improved the *AUC* performance value. Therefore, dual HPO is 64% effective in enhancing the productivity of SVP models.

RQ 2: Is dual HPO better than other HPO scenarios?

Wilcoxon signed-rank test results mentioned in Tables 11–13 show cases of significant improvements. To check whether dual HPO is better than other HPO scenarios, we compare the number of instances of significant improvement of dual HPO with others. AdRn & AdRd (case 2) shows 26 significant improvements out of 75 instances, AdRn & AoRd (case 3 of Section 4.2) shows 38 significant improvements, and AdRn & AdRo (case 4 of Section 4.2) results in 34 instances of significant improvement. Dual HPO has 48 instances of improvement; therefore it is better than other HPO scenarios.

RQ 3: How has the degree of class overlapping affected the HPO?

There are 100 instances (consider cases 3–6 of Section 4.2) of HPO comparisons per dataset. Drupal has shown improvement in 61 instances, Moodle in 75 instances, and PHPMyAdmin in 10 instances. HPO depends on data complexity measures mentioned in Section 3.6. If the overlap among the classes is low, then HPO performs efficiently. As per the $F1$ values for Drupal, Moodle, and PHPMyAdmin in Table 4, PHPMyAdmin has the highest overlap, and Moodle has the lowest. Therefore, for PHPMyAdmin, HPO has not improved

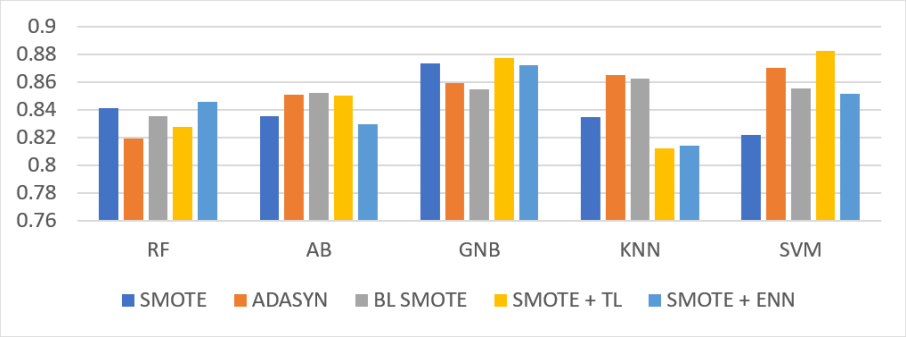

Figure 8. Comparison of resampling techniques in Drupal dataset

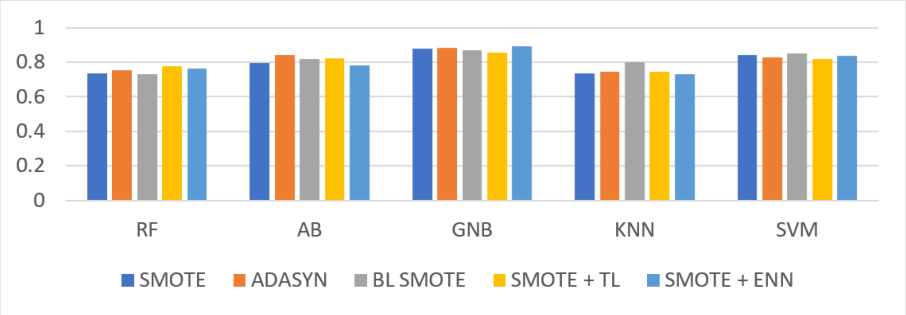

Figure 9. Comparison of resampling techniques in Moodle dataset

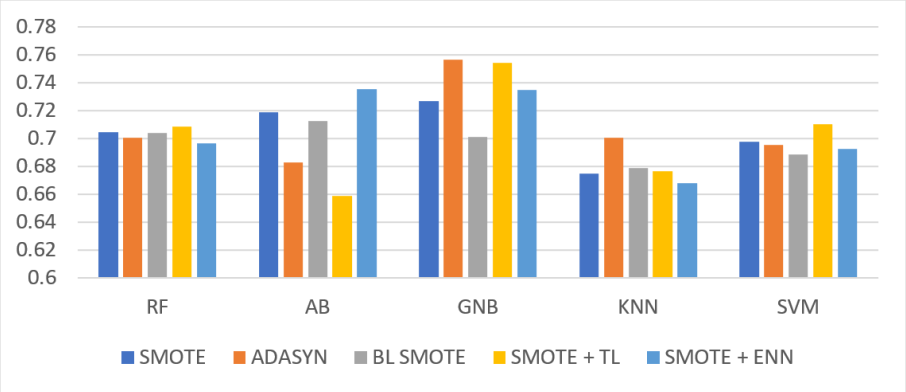

Figure 10. Comparison of resampling techniques in PHPMyAdmin dataset

the performance and, for Moodle, which is highly imbalanced, HPO has worked well. For Drupal, HPO has given mixed results.

RQ 4: Which resampling technique has performed the best?

Figures 8–10 show the highest *AUC* value among four scenarios (AdRd, AoRd, AdRo, and AoRo) resampling techniques for each machine learning algorithm.

It has been observed that there is not one technique that performs well for all algorithms and all datasets. Different machine learning algorithms and datasets have distinct highest-performing resamplers. For instance, in the case of the Drupal dataset, SMOTE + ENN has performed highest in RF, BL SMOTE in AB, SMOTE + TL in GNB and SVM, and finally Adasyn in KNN. For Moodle, SMOTE + TL in RF, Adasyn in AB, SMOTE + ENN in GNB, BL SMOTE in KNN and SVM. For PHPMyAdmin, SMOTE + TL in RF and SVM, SMOTE + ENN in AB, Adasyn in GNB and KNN.

6. Discussions

This section discusses the findings of our study. Imbalanced datasets and lack of hyperparameter tuning may impact the productivity of SVP models. HPO and resampling techniques are known to improve the performance of the prediction models. The current study aims to analyze the effect of dual HPO in software vulnerability prediction. By dual HPO, it means that hyperparameters of both machine learning algorithms and resampling techniques are optimized. We have not only checked the effect of dual HPO but additionally, analyzed whether dual HPO is better than single HPO scenarios where only one of the two factors (machine learning and resampling) is optimized. Furthermore, the inclusion of data complexity measures helps in finding why HPO is not producing results for the classes with high overlap.

Our study will help researchers in improving the performance of their work on prediction models and provide open research for multi-objective optimization, text-mining-based SVP models, cost and time complexity measures, and deep learning approaches.

7. Threats to validity

The current paper deals with the following threats to validity:

- Construct validity: The datasets used in this paper are PHP web application projects. This paper has used a metrics-based dataset to carry out the experiments. If a different type of feature such as text-mining is considered, then the results may vary.
- Internal validity: The selection of machine learning algorithms is based on past research performed. We have confined our experiments to five machine-learning algorithms. Also, the resampling techniques used are either oversampling or a combination of oversampling and under-sampling. Under-sampling techniques are not used because information loss may occur. The hyperparameter search spaces are taken from past research, and results may vary for considering different search spaces. The paper takes care of optimizing the single-objective function which works on increasing the value of the particular metric AUC in our case, a multi-objective function is for the future scope.
- External validity: Generalization of the work in an empirical study is always limited, and it is difficult to conclude. The generalizability of our results is out of the scope of the current study. The performance varies for different programming languages as the features and granularity levels may be different. The study aims to see the impact of dual HPO on various machine learners and to check whether it improves the efficiency of metrics-based SVP models. It uses the Wilcoxon signed-rank test with Bonferroni correction for performing significant comparisons. This test is based on the data sample in hand.
- Conclusion validity: AUC is the performance metric used in the current study. There exist other parameters for the evaluation of imbalanced datasets such as Geometric mean, Matthews's correlation coefficient (MCC), etc. HPO comes with time and cost overhead. We have kept these overheads out of the scope of this paper.

8. Conclusions and future scope

The current study gives an insight into HPO in the machine learning area. Previous studies have used HPO in bug prediction, defect prediction, and even vulnerability prediction

models. In this work, we have analyzed the effectiveness of dual HPO on the capability of machine learners in software vulnerability prediction. Further, whether dual HPO performs better than other HPO scenarios is examined. In addition, it is further studied why HPO performance is degraded for some datasets, i.e., has class overlapping affected the ability of HPO in improving the efficacy of machine learners. The study proposed the experimental methodology based on the python framework "Optuna" that evaluates the six HPO scenarios depicted in Table 1. The paper uses five machine learning algorithms and five resampling techniques for three open-source software vulnerability datasets Drupal, Moodle, and PHPMyAdmin. The best hyperparameters are found for learners and resamplers to optimize the SVP model. In addition to this, the Wilcoxon signed-rank test with Bonferroni correction is applied for statistical comparison to know which HPO scenario has performed significantly. The experimental results are concluded as:

- The results state that dual HPO has shown 64% effectiveness in amplifying the efficiency of SVP models.
- Dual HPO shows 64% effectiveness whereas HPO when applied on machine learners shows 51% and HPO when applied on resamplers obtains 45.33% effectiveness. Therefore, it can be concluded that Dual HPO performs better than other HPO scenarios. We have not compared with the scenarios that do not involve resampling. Although AoRn has shown 9 significant improvements out of 15, we cannot compare them as they may be biased since applied to unbalanced datasets.
- The efficiency of HPO is affected by class overlapping. One of the data complexity measures; is the maximum Fisher's discriminant ratio ($F1$) which calculates the overlap among the classes. The datasets with high overlapping classes result in the poor performance of HPO. PHPMyAdmin is highly overlapped resulting in 10 improvements out of 100; therefore HPO has not performed well for it. Moodle being low overlapped gives the maximum significant results for HPO 75 out of 100.
- Resampling Techniques are analyzed and it is found that they perform differently on distinct datasets and with different learners. Hence, we cannot find the best resampler that fits all the datasets and machine learners.

Future work emphasizes the use of HPO scenarios for text-based datasets. We can consider time and cost complexity measures with more data complexity measures in the future. The impact of HPO on deep learning approaches can be studied. Other programming languages can be explored. Furthermore, the multi-objective problem can be explored.

References

[1] S.M. Ghaffarian and H.R. Shahriari, "Software vulnerability analysis and discovery using machine-learning and data-mining techniques: A survey," *ACM Computing Surveys (CSUR)*, Vol. 50, No. 4, 2017, pp. 1–36.

[2] W.R.J. Freitez, A. Mammar, and A.R. Cavalli, "Software vulnerabilities, prevention and detection methods: A review," in *Security in Model Driven Architecture (SEC-MDA)*, A. Bagnato, Ed., 2009, pp. 6–13.

[3] A. Kaya, A.S. Keceli, C. Catal, and B. Tekinerdogan, "The impact of feature types, classifiers, and data balancing techniques on software vulnerability prediction models," *Journal of Software: Evolution and Process*, Vol. 31, No. 9, 2019, p. e2164.

[4] J. Morgenthaler and J. Penix, "Software development tools using static analysis to find bugs," *Development*, 2008.

[5] B. Arkin, S. Stender, and G. McGraw, "Software penetration testing," *IEEE Security and Privacy*, Vol. 3, No. 1, 2005, pp. 84–87.

[6] P. Godefroid, "Random testing for security: Blackbox vs. whitebox fuzzing," in *Proceedings of the 2nd International Workshop on Random Testing: Co-Located With the 22nd IEEE/ACM International Conference on Automated Software Engineering (ASE 2007)*, 2007, pp. 1–1.
[7] D. Evans and D. Larochelle, "Improving security using extensible lightweight static analysis," *IEEE Software*, Vol. 19, No. 1, 2002, pp. 42–51.
[8] M. Fagan, "Design and code inspections to reduce errors in program development," in *Software Pioneers*, M. Broy and E. Denert, Eds. Springer, 2002, pp. 575–607.
[9] H. Shahriar and M. Zulkernine, "Mitigating program security vulnerabilities: Approaches and challenges," *ACM Computing Surveys (CSUR)*, Vol. 44, No. 3, 2012, pp. 1–46.
[10] M. Jimenez, R. Rwemalika, M. Papadakis, F. Sarro, Y. Le Traon et al., "The importance of accounting for real-world labelling when predicting software vulnerabilities," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019, pp. 695–705.
[11] Y. Shin and L. Williams, "Can traditional fault prediction models be used for vulnerability prediction?" *Empirical Software Engineering*, Vol. 18, No. 1, 2013, pp. 25–59.
[12] T. Zimmermann, N. Nagappan, and L. Williams, "Searching for a needle in a haystack: Predicting security vulnerabilities for windows vista," in *Third International Conference on Software Testing, Verification and Validation*. IEEE, 2010, pp. 421–428.
[13] H. Alves, B. Fonseca, and N. Antunes, "Experimenting machine learning techniques to predict vulnerabilities," in *Seventh Latin-American Symposium on Dependable Computing (LADC)*. IEEE, 2016, pp. 151–156.
[14] W. Almutairi and R. Janicki, "On relationships between imbalance and overlapping of datasets," in *Proceedings of 35th International Conference on Computers and Their Applications*, 2020, pp. 141–150.
[15] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Transactions on Reliability*, Vol. 62, No. 2, 2013, pp. 434–443.
[16] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap," *Procedia Computer Science*, Vol. 176, 2020, pp. 420–429.
[17] K. Borowska and J. Stepaniuk, "Imbalanced data classification: A novel re-sampling approach combining versatile improved SMOTE and rough sets," in *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, 2016, pp. 31–42.
[18] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue et al., "Learning from class--imbalanced data: Review of methods and applications," *Expert Systems With Applications*, Vol. 73, 2017, pp. 220–239.
[19] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, Vol. 465, 2018, pp. 1–20.
[20] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 40, No. 1, 2009, pp. 185–197.
[21] C. Tantithamthavorn, S. McIntosh, A.E. Hassan, and K. Matsumoto, "Automated parameter optimization of classification techniques for defect prediction models," in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 321–332.
[22] J.N. Van Rijn and F. Hutter, "Hyperparameter importance across datasets," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 2367–2376.
[23] H.J. Weerts, A.C. Mueller, and J. Vanschoren, "Importance of tuning hyperparameters of machine learning algorithms," *arXiv preprint arXiv:2007.07588*, 2020.
[24] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, Vol. 415, 2020, pp. 295–316. [Online]. https://www.sciencedirect.com/science/article/pii/S0925231220311693
[25] R. Shu, T. Xia, L. Williams, and T. Menzies, "Better security bug report classification via hyperparameter optimization," *arXiv preprint arXiv:1905.06872*, 2019.

[26] J. Kong, W. Kowalczyk, D.A. Nguyen, T. Bäck, and S. Menzel, "Hyperparameter optimisation for improving classification under class imbalance," in *Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2019, pp. 3072–3078.

[27] R. Shu, T. Xia, J. Chen, L. Williams, and T. Menzies, "How to better distinguish security bug reports (using dual hyperparameter optimization)," *Empirical Software Engineering*, Vol. 26, No. 3, 2021, pp. 1–37.

[28] A. Agrawal, W. Fu, D. Chen, X. Shen, and T. Menzies, "How to "dodge" complex software analytics," *IEEE Transactions on Software Engineering*, Vol. 47, No. 10, 2019, pp. 2182–2194.

[29] A. Agrawal, X. Yang, R. Agrawal, R. Yedida, X. Shen et al., "Simpler hyperparameter optimization for software analytics: Why, how, when," *IEEE Transactions on Software Engineering*, Vol. 48, No. 8, 2021, pp. 2939–2954.

[30] J. Walden, J. Stuckman, and R. Scandariato, "Predicting vulnerable components: Software metrics vs text mining," in *25th International Symposium on Software Reliability Engineering*. IEEE, 2014, pp. 23–33.

[31] J. Stuckman, J. Walden, and R. Scandariato, "The effect of dimensionality reduction on software vulnerability prediction models," *IEEE Transactions on Reliability*, Vol. 66, No. 1, 2016, pp. 17–37.

[32] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," 2015. [Online]. https://arxiv.org/abs/1502.02127

[33] P.K. Kudjo, S.B. Aformaley, S. Mensah, and J. Chen, "The significant effect of parameter tuning on software vulnerability prediction models," in *19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. IEEE, 2019, pp. 526–527.

[34] E. Sara, C. Laila, and I. Ali, "The impact of SMOTE and grid search on maintainability prediction models," in *16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2019, pp. 1–8.

[35] H. Osman, M. Ghafari, and O. Nierstrasz, "Hyperparameter optimization to improve bug prediction accuracy," in *Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE)*. IEEE, 2017, pp. 33–38.

[36] V.H. Barella, L.P. Garcia, M.P. de Souto, A.C. Lorena, and A. de Carvalho, "Data complexity measures for imbalanced classification tasks," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[37] T.K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 3, 2002, pp. 289–300.

[38] J.M. Sotoca, J. Sánchez, and R.A. Mollineda, "A review of data complexity measures and their applicability to pattern classification problems," in *Actas del III Taller Nacional de Mineria de Datos y Aprendizaje*, R. Ruiz, J.C. Riquelme, and J.S. Aguilar-Ruiz, Eds., 2005, pp. 77–83.

[39] A.C. Lorena, L.P. Garcia, J. Lehmann, M.C. Souto, and T.K. Ho, "How complex is your classification problem? A survey on measuring classification complexity," *ACM Computing Surveys (CSUR)*, Vol. 52, No. 5, 2019, pp. 1–34.

[40] Y. Zhang, D. Lo, X. Xia, B. Xu, J. Sun et al., "Combining software metrics and text features for vulnerable file prediction," in *20th International Conference on Engineering of Complex Computer Systems (ICECCS)*, 2015, pp. 40–49.

[41] I. Abunadi and M. Alenezi, "An empirical investigation of security vulnerabilities within web applications," *Journal of Universal Computer Science*, Vol. 22, 2016, pp. 537–551.

[42] M.N. Khalid, H. Farooq, M. Iqbal, M.T. Alam, and K. Rasheed, "Predicting web vulnerabilities in web applications based on machine learning," in *Intelligent Technologies and Applications*. Singapore: Springer Singapore, 2019, pp. 473–484.

[43] C. Catal, A. Akbulut, E. Ekenoglu, and M. Alemdaroglu, "Development of a software vulnerability prediction web service based on artificial neural networks," in *Trends and Applications in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing, 2017, pp. 59–67.

[44] D. Bassi and H. Singh, "Optimizing hyperparameters for improvement in software vulnerability prediction models," in *Advances in Distributed Computing and Machine Learning*, R.R. Rout,

S.K. Ghosh, P.K. Jana, A.K. Tripathy, J.P. Sahoo et al., Eds. Singapore: Springer Nature, 2022, pp. 533–544.

[45] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie et al., "RFRSF: Employee turnover prediction based on random forests and survival analysis," in *Web Information Systems Engineering – WISE 2020*, Z. Huang, W. Beek, H. Wang, R. Zhou, and Y. Zhang, Eds. Cham: Springer International Publishing, 2020, pp. 503–515.

[46] R.E. Schapire, *The Boosting Approach to Machine Learning: An Overview*. New York, NY: Springer New York, 2003, pp. 149–171.

[47] R. Meir and G. Rätsch, *An Introduction to Boosting and Leveraging*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 118–183.

[48] R. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, Vol. 37, 1999, pp. 297–336.

[49] H. Zhang, "The optimality of Naive Bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, V. Barr and Z. Markov, Eds. AAAI Press, 2004.

[50] M. Martinez-Arroyo and L.E. Sucar, "Learning an optimal naive bayes classifier," in *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3. IEEE, 2006, pp. 1236–1239.

[51] J.D.M. Rennie, L. Shih, J. Teevan, and D.R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 2003, pp. 616–623.

[52] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, 2011, pp. 1–27.

[53] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V.N. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9 (NIPS)*, M. Mozer, M. Jordan, and T. Petsche, Eds., 1996.

[54] X. Wang, J. Yang, X. Teng, and N. Peng, "Fuzzy-rough set based nearest neighbor clustering classification algorithm," in *Fuzzy Systems and Knowledge Discovery*, L. Wang and Y. Jin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 370–373.

[55] W. Zuo, D. Zhang, and K. Wang, "On kernel difference-weighted k-nearest neighbor classification," *Pattern Analysis and Applications*, Vol. 11, No. 3, 2008, pp. 247–257.

[56] M. Santos, J. Soares, P. Henriques Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches," *IEEE Computational Intelligence Magazine*, Vol. 13, 2018, pp. 59–76.

[57] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, Vol. 16, 2002, pp. 321–357.

[58] H. He, Y. Bai, E.A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[59] H. Han, W.Y. Wang, and B.H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, D.S. Huang, X.P. Zhang, and G.B. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887.

[60] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. abs/1907.10902, 2019. [Online]. http://arxiv.org/abs/1907.10902

[61] V. López, A. Fernández, J.G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Systems with Applications*, Vol. 39, No. 7, 2012, pp. 6585–6608.

[62] J. Huang and C.X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 3, 2005, pp. 299–310.

[63] K. Sultana, B. Williams, and A. Bosu, "A comparison of nano-patterns vs. software metrics in vulnerability prediction," in *Proceedings – 25th Asia-Pacific Software Engineering Conference, APSEC 2018*. IEEE Computer Society, 2018, pp. 355–364.

[64] A.K. Tanwani and M. Farooq, "Classification potential vs. classification accuracy: A comprehensive study of evolutionary algorithms with biomedical datasets," in *Learning Classifier Systems*, J. Bacardit, W. Browne, J. Drugowitsch, E. Bernadó-Mansilla, and M.V. Butz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 127–144.
[65] S.S. Rathore and S. Kumar, "An empirical study of ensemble techniques for software fault prediction," *Applied Intelligence*, Vol. 51, No. 6, 2021, pp. 3615–3644.
[66] D. Tomar and S. Agarwal, "Prediction of defective software modules using class imbalance learning," *Applied Computational Intelligence and Soft Computing*, Vol. 2016, 2016, pp. 1–12.
[67] A. Kaur and K. Kaur, "Statistical comparison of modelling methods for software maintainability prediction," *International Journal of Software Engineering and Knowledge Engineering*, Vol. 23, 2013.
[68] E.W. Weisstein, *Bonferroni Correction*, 2004. [Online]. https://mathworld.wolfram.com/
[69] P. Sedgwick, "Multiple significance tests: The Bonferroni correction," *BMJ (online)*, Vol. 344, 2012, pp. e509–e509.

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230103, DOI: 10.37190/e-Inf230103

Computer Game Scenario Representation: A Systematic Mapping Study

Maria-Eleni Paschali*, Ioannis Stamelos*
*Department of Informatics, Aristotle University of Thessaloniki, Greece
mpaschali@csd.auth.gr, stamelos@csd.auth.gr

Abstract

Background: Game scenario is an important factor for achieving player enjoyment; consisting a key business success factor. Additionally, the production of early design artifacts is crucial for the success of the development process. However, representing scenarios is a non-trivial task: (a) multiple aspects of the game need to be visualized; and (b) there is a plethora of representation approaches, out of which the game designer needs to select from.
Aim: The goal of this work is to provide a panorama of the current scenario representation approaches, to aid game engineers in selecting the most fitting scenario representation approach and understand the existing designing options.
Method: We have performed a Systematic Mapping Study, using 4 digital libraries, since the main goal can be achieved through study classification. By following an established search and filtering process, we have identified 717 articles, and analyzed in detail 95.
Results: Diagrams are the most common generic approach to represent scenario; Game story is the most usual part of the scenario being represented; Characters are the most common component; and Transitions are the most usual connectors.
Conclusion: Researchers may get useful information for empirically investigating several game engineering aspects; whereas game engineers can efficiently select the most fitting approach.

Keywords: systematic reviews and mapping studies, software architectures and design

1. Introduction

The rapid technological developments have led to a massive increase of the size of the whole spectrum of digital activities, causing multiple changes to everyday routines of humans. One prolific example of this rise, is the growth and popularity of computer games [1]. Nowadays, computer games are an important part of the entertainment for both children and adults. Children choose to devote more and more time to this kind of entertainment by sacrificing other activities [2]. The large spread of digital games is proven by the enormous availability of game titles and the fact that game industry could compare to that of "*Hollywood industry*" [3]. The games have the ability of attracting the interest and the full attention of gamers, "*making*" 78% of American teenagers to play computer games [2]. According Statista, the revenue from computer and console games is expected to reach $240 billion

Submitted: 18 Jul. 2022; Revised: 3 Nov. 2022; Accepted: 20 Nov. 2022; Available online: 6 Dec. 2022

in 2026 from $175 billion in 2022[1]. Although, games form a special kind of software (e.g., the most important factor for their success is user enjoyment and usability, rather than functional correctness and suitability), they still obey to all software engineering principles. In other words, games despite being a collection of graphics, animations, sounds, code, etc. the core product is a collection of code artifacts that needs to be specified (what the game will do), designed (game architecture), implemented (actual coding of classes) and tested (functional, usability, etc.), like any other software solution. Based on this, game analysis and design (e.g., scenario, character definition, etc.) are of paramount importance for the successful implementation of the games, as well as their success.

According to Ham and Lee [4] and Paschali et al. [5] there are seven high-level characteristics that lead to gamers' satisfaction, engaging them to game playing; namely: Scenario, Graphics, Speed, Sound, Control, Characters, and Community. Both papers conclude that Character Solidness, Scenario and Sound are highlighted as the most crucial deciders on if a game will be successful or not. By considering that character building (i.e., definition, relationship, interactions, and so on.) are part of the game scenarios, ***scenarios can be promoted to the most important factor for game success***, since both studies point to this direction. In this work scenario we define the description of the game in terms of plot, world, rules, characters, interaction, and any other element that is required to describe and specify a game. Given the complexity and dynamic nature of scenarios, their representation in game design documents is far from trivial [6]. However, being able to represent a scenario properly at an early design stage is of paramount importance, since: (a) it acts as a ***communication vehicle*** among different development stakeholders, such as: designers, developers, scenario artists, graphic experts, and so on.; and (b) it acts as an early piece of documentation that can be ***easily perceived by end-users and be an artifact for early testing***. One important parameter that needs to be considered before deciding the representation approach, is the game genre (e.g., Action, Adventure, Arcade, Realtime strategy, God Games, Roleplaying, Shooter, Simulations, Sport, Strategy, and so on.). At the design phase, according to the genre of the game, the way to depict the scenario is chosen: along with its components, connectors and so on. More specifically, according to the genre of the games the game rules, world, and mechanics differ substantially. However, in academic literature there are only limited studies that focus on this aspect of game design, despite the fact that game engineering literature is continuously growing [7] . Therefore, interested parties (researchers or practitioners) need to read various articles, only superficially connected to the aspect, and gain unconnected and synthesized knowledge.

Given the above, in this paper we present the first (to the best of our knowledge) mapping study to provide a complete panorama of the research state-of-the-art on scenario representation. In particular, we explore the representation methods, as well as the components and the connectors used in these representations. In Section 2, we present related work including secondary studies in the field of game engineering; in Section 3 the study design; whereas in Section 4 we present and in Section 5 discuss the results. Finally, threats to validity are presented in Section 6; we conclude the paper in Section 7.

[1]https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide

2. Background Information

In this section, we present the necessary background information for facilitating the understanding of this article. We note that we have not identified any study that is directly comparable (direct related work), i.e., a secondary study on game scenarios. We need to note that indirectly related secondary studies (such as: secondary studies on game engineering (e.g., [7], or visualization techniques [8]) are not discussed, since we cannot contrast our findings to them. Therefore, we are focusing on background information on game scenarios.

Game Scenario Importance & Scenario Design: According to Ham and Lee [4] and Paschali et al. [5], game scenario is one of the most important features that lead to player satisfaction. Silva (2019) also emphasized how important fun is in serious games for players to want to continue playing and learning as a result [9]. Zyda [10], in the same reasoning, argued that a serious game must first be fun [10]. Zemliansky and Wilcox [11] also mentioned the need for a balance between art and game design to achieve learning while still creating an enjoyable user experience [11]. Many researches aimed at the narrative structure of a scenario. According to Partlan et al. [12] the automated representation of interactive narrative consists of four types of related graphs: (a) the scene graph, which represents how the scenes connect to each other; (b) the layout graph representing the physical placement of objects in the visual environment; (c) the script graph contains the code to operate the scene's gameplay logic; and (d) the interaction map using static graph analysis [12]. At same path, Segel & Heer [13] after analyzing 58 collected examples from online journals, graphic designs, comics, business, art, and visualization research, they identified distinct genres of visualization using narrative structures such as the martini glass, interactive slideshow, and drill-down story [13]. On the other hand, some developers use flowchart for designing game scenario [14], providing an interface that is easier to adopt, use, debug and tune [15]. A tool that we also met is the Code City that is used to visualize cities in games and gives a great variety of opportunities such as interactivity, scalability, navigation and completeness [16].

According to Fabricatore [17] the gamers focus on: (a) what the player can do; and (b) what other entities can do, in response to player's actions (i.e., how the game responds to player's decisions, this would happen with usage of game mechanics [17]. The importance of interaction through game mechanics was also highlighted [18]. Game content, by focusing on the Procedure Content Generation areas, has six layers [19]: (a) game bits, which are elementary units of game content; (b) game space, the environment in which the game takes place; (c) game systems, to generate or simulate parts of a game; (d) game scenarios the way and order in which game events unfold; (e) game design which consist of goals and rules; and (f) derived content is created as a side-product of the game world. Finally, game design is composed of [20]: (a) Features; (b) Gameplay rules; (c) Learning contents; (d) Interface; and (e) Game Levels. Specifically for scenario design the authors identified that the key elements are: (a) blocking tissues; (b) vital tissues, and (c) targets which are modeled with boxes.

The development of games is characterized by a lack of formalization compared to software development. Park and Park [21] proposed a graph-based representation of game scenarios, a combination of Event graph, State graph, and Action graph forms to eliminate anomalies of game flow design and increase the better communication of game designer and the game programmer. The use of design patterns was suggested by Killi [22], who proposed 6 categories of patterns for serious games: (a) Integration Patterns; (b) Cognitive

Patterns; (c) Presentation Patterns; (d) Social Interaction; (e) Teaching Patterns; and (f) Engagement Patterns [22]. Additionally, Amory [23] proposed a new more detailed model, GOMII, as the new version of GOM which in order to support parameters that educational computer games should have: relevant, explorative, emotive, engaging, include a variety of challenges, democratic, include computer tools that promote dialogue, gender-sensitive, provide non-negotiable results, include correct role models [23]. All these parameters made this model not only a tool for supporting learning process but also an evaluating mechanism of computers' usage into classrooms. Finally, the Game content, focusing on Process Content Creation areas, has six levels according to Hendrikx et al. [19]: (a) game bits, which are elementary units of game content; (b) game space, the environment in which the game takes place game; (c) game systems, to create or simulate parts of a game; (d) game scenarios in the manner and order in which game events unfold; (e) game design consisting of goals and rules and (f) the resulting content is created as a side-product of the game world.

3. Methods

This section presents the protocol of the systematic mapping study. A protocol constitutes a pre-determined plan that specifies the research questions and how the mapping study has been conducted. Our protocol is presented according to the guidelines suggested by Peterson et al. [24], whereas the reporting of the secondary study is based on the SEGRESS guidelines [25] – the checklist is presented in Appendix B.

3.1. Objectives and research questions

According to Goal-Question-Metrics (GQM) format, we set the main goal of the study which is to analyse the representation methods of game scenarios. To fulfil this goal, we have set the followed questions, to study scenario representation from three perspectives: (a) the ways of their representation; (b) the parts of scenarios that are represented; and (c) the components of the scenario and their connection and we set the followed questions.

RQ_1: Which are the most common methods in the academic literature for representing computer game scenarios?

As the scenarios are complex and dynamic, there is a need to find an appropriate way to depict them in game design documents. RQ_1 is related to the identification of the methods for game scenario representation. This question is answered at two levels, since we build a 2-level classification schema. First, we identify the *Generic Representation Type (GRT)*, and in the next step we specify (whenever relevant), a more *Specific Representation Type (SRT)* – e.g., as proposed by the Unified Modelling Language (UML). Examples of SRTs are state-machine, flow chart, activity diagram, class diagram or sequence diagram. For each generic representation type, we explain in detail the meaning, and then we present all the pertaining SRTs.

RQ_2: What parts of the scenario are captured by the representation methods?

RQ_2 is related to the exact parts of the scenario that are depicted in each representation method identified in RQ_1. To answer this question, we separate the scenario into three parts. First, Game Story, which presents the flow of events in the game and captures aspects such as player navigation, the options of the player and so on. Second, Game World that represents the visual elements of the game including the description of the world locations

along with the characters and objects they include. Finally, Game Rules that correspond to the mechanics that control the flow of the game. A rule can be related to an action of the player in conjunction with the state of the world the characters and so on. The motivation of this question is to decompose the scenario representation approach selection problem to smaller ones, so as to be more a manageable. The answer to this question will expand the previous schema, by noting the parts of the scenario that can be represented by each GRT or SRT. Based on this, we suggest combinations of representation methods that are able to capture all parts of a scenario.

RQ_3: What elements (components and connectors) are used in the aforementioned representation approach?

In RQ_3 we focus on the different representation methods, and we seek mappings between representation elements (i.e., components and connectors) to game elements. For example, as component we can consider the character of a game and as connector the actions of the character. The reason for asking this question is for creating a checklist of the elements that need to be considered for every part and guiding the scenario design process in a more organized way.

3.2. Search process

Our search strategy has been developed, based on the goal of the study and the set research questions. Based on these, we opted for performing a mapping study, rather than a systematic literature review, since: (a) the topic is broad; (b) we aimed at a generic overview of the topic; (c) the main goal of the study is to provide a classification of scenario representation approaches. Based on the above, we performed an automated search through the advanced search functions of four well-known Digital Libraries (DL): (a) IEEE Digital Library, (b) ACM Digital Library, (c) ScienceDirect; and (d) Scopus. We opted for searching in DL instead of narrowing the search space to specific venues, since we are not interested in specific communities or publication sources (e.g., only software engineering or only graphics). As a first step we applied the search string to the abstract of primary studies in Q1 of 2021, to return all the papers that are relevant to game scenario. The search string is described below:

[**Abstract**: game or gaming] AND
[**Abstract**: visualization or design or depiction or representation] AND
[**Abstract**: scenario]

The decision to apply the search string to the Abstract has been made by piloting that the same search string on the Title misses various important sand highly relevant studies. The main reason for this is that many authors use in the title “Game” or “Scenario”, or a specific representation approach (e.g., “Flow Chart”), rather than the terms of the 2nd part of the string (visualization or design or depiction or representation). The alternative to this would have been to search the title for games, and add the scenario representation approach as an inclusion criterion. However, this option would return an unmanageable amount of candidate primary studies (as you will later see the exclusion rate was quite high even for the narrower search string). Another alternative would have been to build a more specific search string that would return less articles that would be more relevant (e.g., by including the expected representation approaches). However, this would have biased our results, since the data collection would not be open ended, and there would be a higher

chance of missing papers. Given the two corner solutions (too generic or too specific) search string, we have opted for the middle path which would not bias the results, but would provide a large, but manageable corpus of papers for the IC/EC process. We selected each word and its synonym in order to eliminate the possibility of losing relative articles.

After retrieving the first dataset we defined the Inclusion Criteria (IC) and Exclusion Criteria (EC). A primary study has been selected for inclusion, if it satisfied the first IC and one or more of the rest ICs, whereas it has been excluded from our study, if it satisfied one or more ECs. The inclusion criteria of our systematic mapping are:

– IC1: The primary study is applied in computer games for instance, primary studies referring to "traditional" games, without using a computing system have been excluded;
– IC2: The primary study contains a method of game scenario's representation; we need a paper to present a scenario of a computer game, that is presented using a specific method – ranging from textual descriptions to graphical representations.
– IC3: The primary study presents building blocks that consist the scenario; the scenario is not presented as a complete block, but is separated into parts.

The exclusion criteria of our systematic mapping are:

– EC1: The primary study is written in a language other than English.
– EC2: The primary study is an editorial, keynote, biography, opinion, tutorial, workshop summary report, progress report, poster, or panel.
– EC3: The primary study contains only a part of scenario and not the whole scenario; for example, in a ping pong game only the move of the ball is depicted.

Every article selection phase was handled by one member of the team and possible difficulties were resolved by the other member. For each selected publication venue, we documented the number of papers that were returned from the search and the number of papers finally selected. The main reasons for filtering out papers were: (a) their focus on "real-life" (e.g., original monopoly or physical sport games) and not "computer" games – approximately 40%; or (b) the lack of a specific representation approach – approximately 35%. At the end of the process, we have obtained a dataset of 95 primary studies. An overview of the aforementioned process is provided in Figure 1.

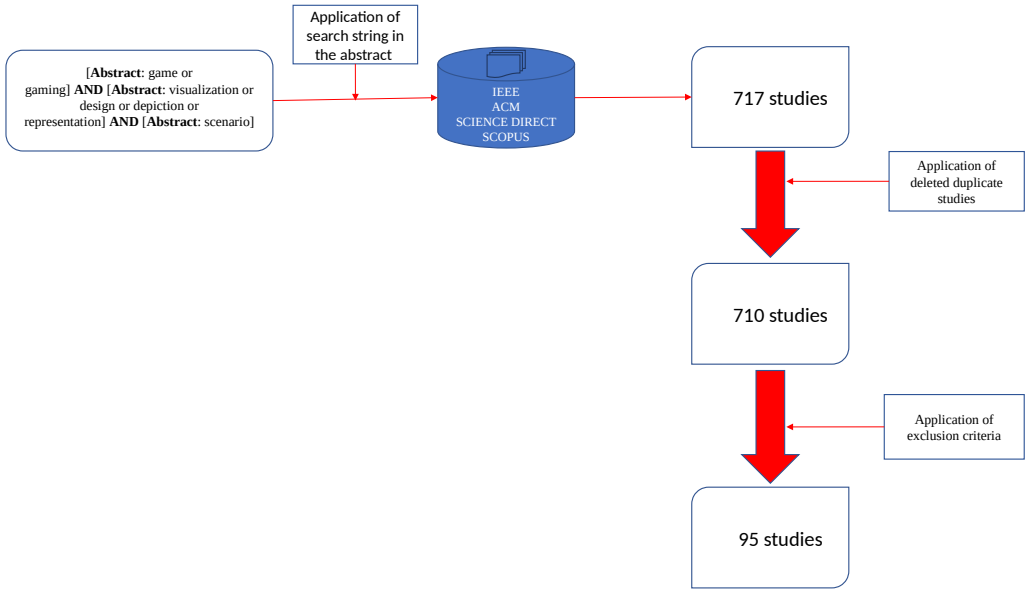

Figure 1. Overview of search process

3.3. Data collection and analysis

During the data collection phase, we collected a set of variables that describe each primary study. Data collection was handled by the first author and possible difficulties were resolved by the second author. For every study, we extracted and assigned values to the following variables:

V1 Title: Records the *title* of the paper.
V2 Author: Records the *list* of authors of the paper.
V3 Year: Records the *publication year* of the paper.
– Type of Paper: Records if the paper is announced/published in a conference or *journal or workshop*.
V4 Publication Venue: Records the name of the corresponding journal or conference.
V5 GRT: Records the *generic scenario representation* approach (e.g., narrative structure, UML)
V6 SRT: Records the *specific scenario representation* approach (e.g., algorithm, Petri-Net, pseudocode)
V7 *Components* of game representation (e.g., characters, dialogs)
V8 *Connectors* of game representation (e.g., link two characters through a dialog)
V9 *Part of the game scenario* that is represented? (e.g., rules, story, game world)

The variables have been selected, based on the set RQs, and they are used to answer them, through 1-to-1 matching, see below. The complete dataset for this study is presented in Appendix A. Appendix A, serves also as the final list of primary studies considered in this work. Due to the large number of scenarios' representation in the literature we performed pre-processing. To group more general categories, we used Open Card Sorting . We have selected to use Open-Card sorting since it is an established method for coding in the literature, it is rather simple and straightforward, and it can be applied by the small number of authors of this research. In particular, we: (a) identified more general categories (e.g., UML generic type) from the scenarios' representation methods in the primary studies; (b) reviewed the methods to find candidates for merging – e.g., we mapped "state machine" as sub-category; and (c) defined the names of the final super-categories and sub-categories. To answer the aforementioned RQs, we chose different ways for presenting the results. More specifically for answering RQ_1 for generic scenario representation approach we present a pie chart and a diagram for combining the generic specification and specific representation approach. For answering RQ_2 we used a pie chart for presenting the frequency of parts of game scenario, a bar chart for parts of scenario represented by generic scenario representation approach and a Venn diagram for representing the parts of scenario represented by specific scenario representation approach. For answering RQ_3 we used two heatmaps: (a) to specify the frequency of connectors used in different scenario representation approaches; and (b) to present generic scenario representation approach with components

4. Results

In this section, we present the results of this study, organized by research question. Therefore, in Section 4.1, we present the most common ways of depicting scenarios in game development (RQ_1). In Section 4.1, we present the parts of the scenarios captured by the representation method (RQ_2). Finally, in Section 4.3, we present the elements (components

and connectors) that are used in each game scenario representation approach (RQ_3). As a first step in Table 1, we present a descriptive analysis of the dataset.

Table 1. Descriptive analysis of the dataset

Type of publication	Number of published articles
Articles in Journal	29
Book Chapters	1
Articles in Conference	65
Period	**Number of published articles**
2002–2005	4
2006–2010	15
2011–2015	31
2016–2020	40
2020–2022	5

4.1. Scenario approaches (RQ_1)

This section answers RQ_1 regarding the ways of scenario's representation. Scenario representation approaches can be classified into a 2-level schema: the first level for general ways of representation, further specified in the second one. In Figure 2, we present all the approaches that are used for scenario representation, through a pie chart, using different colours and labels to represent the different approaches. The most common way of representation is by ***diagrams*** for the purpose of visualization, e.g., in ; followed by ***narrative*** that textually describes the details of a scenario for example in. The Unified Modelling Language (**UML**) is the third choice of researchers, e.g., [26] – although in some cases it could be classified into diagrams as well. By considering the UML is usually expressed in the form of a diagram, the generic "diagram" accounts to more than 50% of representation methods. We preferred to present it as a separate representation way for explicitness. The main advantage of using generic (or UML) diagrams is the visual representation, which is usually for easily perceived by the human cognition. On the other hand, the main benefit of using narrative is that is a form of representation/description that can be produced and read without any prior computer science knowledge (e.g., sound or visual artists, script writers). The ***pseudocode***

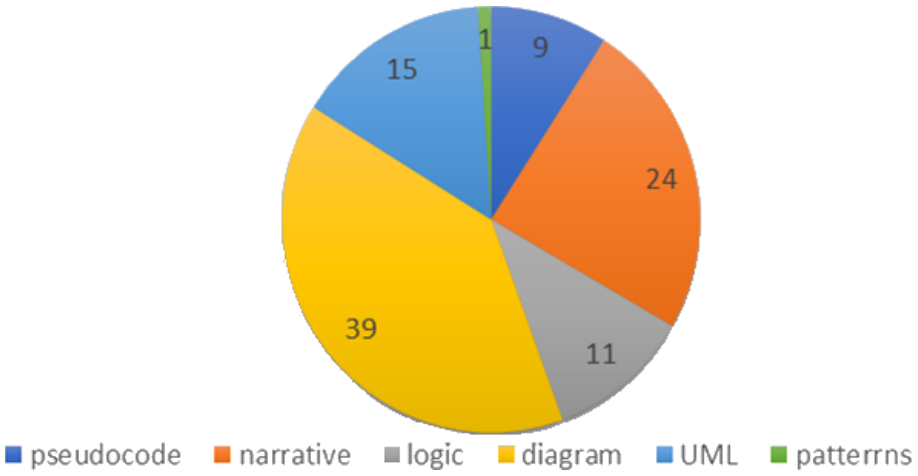

Figure 2. Generic scenario representation approaches

is a program written in "human language" following the programming rules, being used in nine studies for example in [27]. ***Patterns*** and ***logic*** are more rare representation approaches, e.g., [28] and [29].

Next, we focus on specific representation approaches and how they relate to generic ones. From the findings it is clear that the ***state machine diagram*** is by far the most popular specific way for game scenario representation. The second is the ***algorithms***, the third is the ***MAP***. The state machine diagrams with diagram are the most common mapping, e.g., [30], followed by state machine diagram with UML, e.g., [31], algorithms with diagrams, e.g., [32] and flowcharts with diagrams, e.g., are both in the third position with eight appearances and in the fourth position is the pseudocode with algorithms, e.g., [33]. In the fifth position is the logic with algorithms, e.g., in the sixth position is the logic with state machine diagrams. Most cases have one or two occurrences, as shown in Figure 3.

Based on the findings one can observe that there are quite many diverse approaches for describing scenarios. For instance, character models are meant to be used for describing characters, i.e., a very specific element of games; whereas state diagrams can describe a large variety of elements in the design: the state of objects, state of characters, etc. Thus, it is interesting to note that in order to describe a game in perfect detail, a lot and very diverse mechanism are required. A similar finding applies to more high-level aspects of game development, where a teams need various skills (developers, 3D artists, texture artists, animators, sound engineers, writers, etc.) and knowledge of various technologies (programming languages, scripting languages, game engines, 3D editors, etc.). In that sense, we believe that it is reasonable to expect that various and diverse ways of representing scenarios will be needed for designing and representing a game.

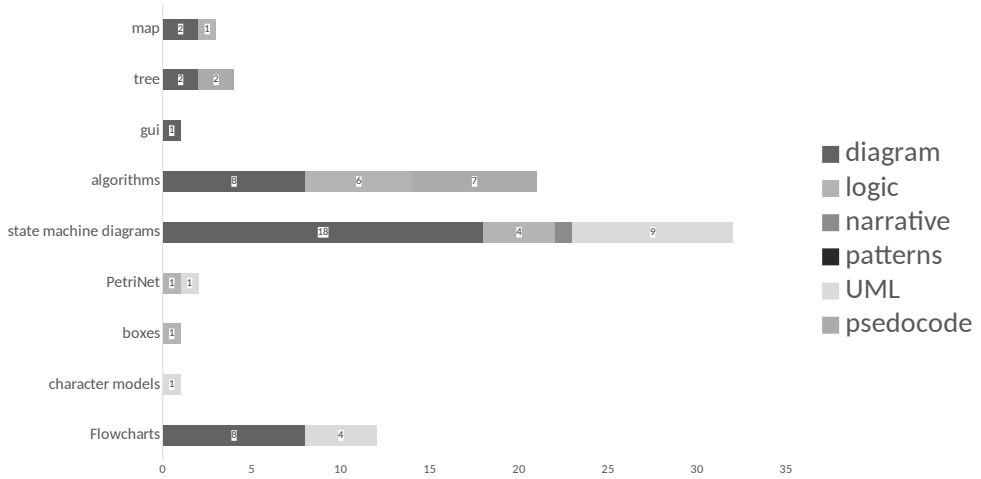

Figure 3. Generic scenario representation approach with specific scenario representation

4.2. Parts of scenario represented (RQ_2)

In Figure 4, we present the mappings of game scenario parts to the generic scenario representation approaches identified in Section 4.1. Game story and game rules are usually represented by diagram [30, 32] followed by narrative, e.g., [34]. Then, game story is presented by UML [14], whereas game rules and game world that are presented by narrative

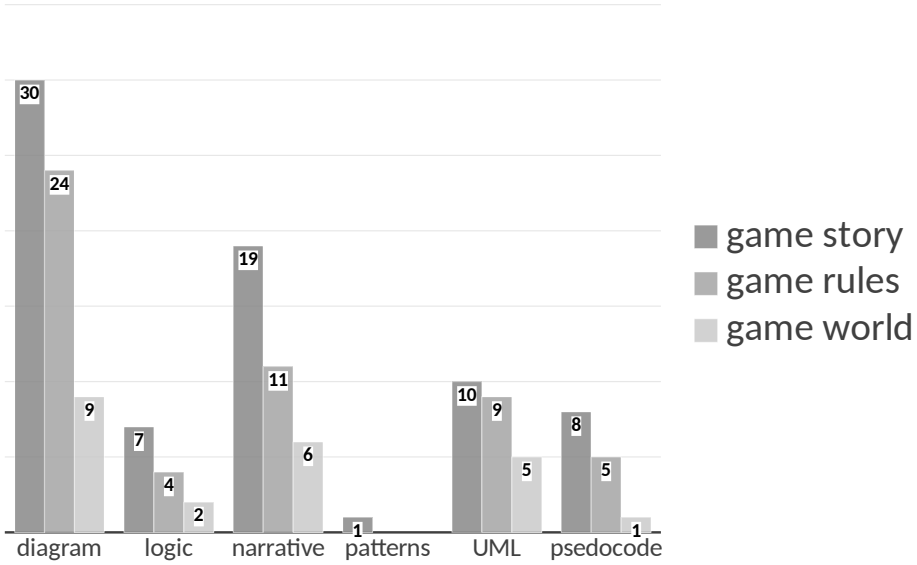

Figure 4. Parts of scenario represented by generic scenario representation approach

Figure 5. Parts of scenario represented by specific scenario representation approach

approaches. Next, in Figure 5 we present the three parts of game scenario parts, mapped to specific representation approaches through Venn diagrams. From the Venn diagram, we can observe that state machine diagrams, algorithms, flowcharts, Petri Networks and Maps can be used for describing all parts of scenarios (story, world and rules), e.g., [35] Game Story and Game World are not simultaneously described by the same specific scenario representation approach, whereas various combined representations between Story and Rules; and Rules and World can be observed.

4.3. Game representation elements/components and connectors (RQ_3)

In this section we present the main elements of scenarios that are depicted through the identified representation approaches. Following the established software architecture terminology, we divide these elements to components and connectors. ***Components*** are the elements that constitute the scenario as it comes up from the description. Below, we provide a list of the components that we identified in the papers:

- Game state: different situations that the game has, e.g., mini-games [27], phases [36, 37].
- Character state: the different states that the characters have, e.g., behaviors [38].
- Time: as an identifier for best player [39], no answer in a question [40], complete the game.
- Characteristics: color [41], accessible, visible, price [42].
- Dialogues: If they are responses of a player, then they are part of game rules. In this case they determine the evolution of the game.
- Animations: Lights, sound, virtual environment, noise.
- Characters: could be also the enemies.

On the other hand, ***Connectors*** are the elements used in order to join elements:

- *Transitions*: one action from one component could cause an action of another component.
- *Actions*: When a component does something.
- *Sequence*: When events happen one after the other.
- *Controls*: When a button is pressed then one action happens

Table 2. Connectors with generic representation approaches

	Sequences	Choices	Flow	Conditions	Navigation	Actions	Transitions	Controls
Diagram	1	1		2	2	11	27	6
Logic	3			2	1	5	1	1
Narrative	2	1	1			11	9	5
Patterns						1		
UML	2	1	2	1		7	8	1
Pseudocode	1					2	7	1

In Table 2, we present the connectors that are used for each representation approach: with dark grey, we denote the most usual connectors used in a scenario representation approach. Based on the results, when the representation approach is a diagram, the most common connector between components are transitions, followed by actions. These connectors are used also in the narrative and UML representation approaches. Finally, transitions are also used in pseudocode and diagram with controls.

In Table 3, through a heatmap we present the percentage rates of the combination among scenario components and general scenario representation approaches. The dark red cells indicate biggest percentages, whereas white the smallest ones. In the following discussion, we exclude representation approaches with small frequency (e.g., patterns). From the results we can observe that diagrams represent at 18% of the cases characters, 12% objects, and at 10% of the cases locations and goals. Also, narrative descriptions focus on goals, scores, objects, characters, locations, etc.

Table 3. Generic scenario representation approach with components

	Diagram	Logic	Narrative	UML	Pseudocode
Locations	10%	12%	9%	14%	3%
Speed	1%	3%	3%	5%	0%
Object	12%	21%	14%	12%	9%
Characters	18%	9%	16%	17%	14%
Character state	2%	6%	1%	3%	3%
Coals	10%	6%	11%	8%	11%
Dialogues	4%	0%	2%	7%	6%
Answers	5%	0%	1%	0%	3%
Questions	4%	0%	2%	0%	3%
Scenes	2%	3%	1%	3%	3%
Levels	3%	0%	3%	0%	9%
Score	5%	3%	10%	0%	6%
Time	5%	0%	5%	3%	3%
Game state	3%	3%	0%	5%	6%
Decisions	8%	6%	1%	3%	11%
Animations	3%	0%	5%	3%	9%
Move	2%	12%	8%	7%	0%
Characteristic	1%	3%	1%	3%	0%
Events	2%	12%	0%	5%	3%

5. Discussion

Comparison to Related Work: In this section, we discuss the main findings of our work, and complement them with existing evidence from previous literature. First, the need for scenario representation is highlight also Partlan et al. [12], who propose an automated representation of interactive storytelling which consists of four types of relational graphs: (a) the scene graph, which represents how scenes are connected to each other; (b) the layout graph representing the physical placement of the objects in the visual environment; (c) the script graph contains the code to operate the game logic of the scene and (d) the interaction map using static graphical analysis [12]. Second, with respect to the use of flow-charts as an important way of scenario representation, we have found various studies that explain in detail how flow charts can be used in scenario design. For instance, Paschali et al. [14] and Tovinkere and Voss [15] provide tools that are ease to adopt, use, debug and tune. Additionally, the Code City tool was used in another study [16] to visualize cities in games and gives a wide variety of opportunities such as interactivity, extensibility, navigation and completeness. The need of connectors is also emphasized in Fabricatore [17]. Players focus on: (a) what the player can do; and (b) what other entities can do in response to the player's actions (i.e., how the game responds to the player's decisions, this would occur using game mechanics. The importance of interaction through game mechanics was also highlighted in Sedig et al. [18].

Synthesis of Results: The findings of the research questions are synthesized in Figure 6; in which, we present an overview of the approaches of game scenario representation. As in most mapping studies our main outcome is a classification schema. The classification has been built based on the raw data of the mapping study. For readability reasons, while developing the classification schema we preferred not to list the primary studies that should be mapped to each edge. A more detailed representation, in a tabular format is presented in Appendix C. According to Nickerson et al. [43], the most common paradigm for building classification schemas for information systems is the three-level indicators model, which is based on both empirical and deductive approaches [43]. By applying this

model, we: (a) examined the objects (i.e., studies), (b) we identified general distinguishing characteristics of the objects, and (c) we grouped their characteristics so as to create our classification schema [43]. Specifically, in step (b) we identified three characteristics that will constitute the three levels of the proposed schema: (a) the 1st level of the schema represents the part of the scenario is depicted; and (b) the 2nd level represents the proposed variables that constitute the three categories of representation method, such as the navigation, the options of a player, the characters, objects, etc.

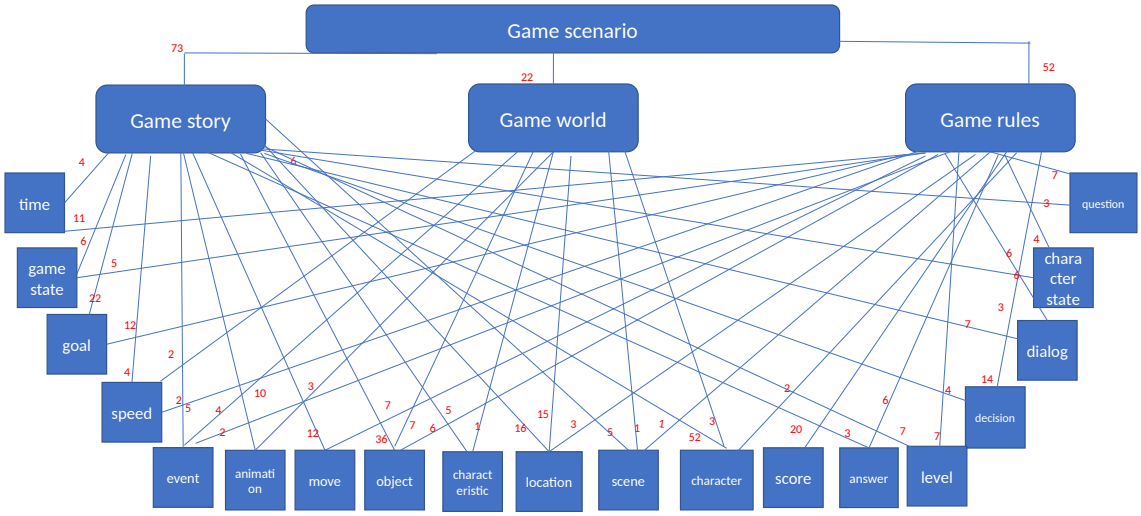

Figure 6. Game scenario representation overview

Implications to Researchers and Practitioners: Based on the classification schema presented in Figure 6, practitioners can: (a) first map the parts of game scenarios to components and connectors, and perform consistency checks – i.e., identify parts of the game scenario that are missing elements descriptions and interactions; (b) second, based on Figure 6, the selected elements can be represented using the most fitting representation approach. In particular, we believe that the following checklist can be used while representing game scenarios to aid practitioners in their scenario design tasks. As a next step, for implementing every item of the checklist, the practitioners can refer to Figure 3 for the most fitting ways of representing each game scenario component. For instance, when the game designer wants to represent game characters (as part of game story), the most frequent means of visualization is through diagrams (4th line of Figure 3); whereas more details can be retrieved from the 52 primary studies mentioned in the last line of the table in Appendix C.

Part A: Have you designed/represented the Game Story?
- Have you explicitly stated the goals of the game?
- Have you explicitly described the objects of the game?
- Have you explicitly described the locations and the levels of the game?
- Have you represented the characters (and their animations) of the game?
- Have you specified the dialogs (Q&As) of the scenes?
- Have you the game state, and the corresponding transitions?

Part B: Have you designed/represented the Game World?
- In the game world, have you specified precise locations?

- Have you mapped locations to game elements (objects, characters)?
- Have you set the locations of scenes and stories?
- Have you set the physics of the world (e.g., speed)?

Part C: Have you designed/represented the Game Rules?

- Have you explicitly specified the rules on when each goal is reached?
- Have you explicitly specified the rules for scoring?
- Have you explicitly specified all the decisions (and possible outcomes) that the user must make?
- Have you guaranteed that the parameters for the rules correspond to game elements (objects, time, characters, moves, levels, dialogues, etc.)?

On the other hand, **researchers** can identify parts of scenario that lack representation approaches, and introduce most fitting ones. Additionally, another interesting future work direction is the provision of tools that not only visualize scenario elements, but also use AI to safeguard the conformance to the aforementioned constraint. Finally, we believe that an interesting future work direction would be the empirical evaluation of the effectiveness and useability of these representation approaches in the game design industry; e.g., organize a workshop that would ask practitioners to represent the same game using different approaches, and later perform focus groups to highlight the pros and cons of each approach.

6. Threats to validity

In this section, we present threats to validity based on the guidelines as supported by Ampatzoglou et al. [44] and [45]. According to Zhou et al. [45] one of the mechanisms of ensuring the level of scientific value in the findings of an SLR is to rigorously assess its validity; in that sense, in this section we identify and report the threats to validity for this study. The classification of threats is performed based on Ampatzoglou et al. [44], since it is a more recent and extensive study.

Study Selection Process: Study selection concerns the first steps of performing the secondary study process, when we had specified the search string to return us the papers related to our subject and filter the most relevant ones for our purposes. To examine the primary studies for inclusion, we had followed a specific protocol based on strict guidelines [46]. The search process has been performed using the search engine of DLs, with specific filters according to our requirements. As the subject is quite general, we had not chosen a broad search string that would lead to an enormous number of papers, so we limited the search space by using quotation marks and searched only the title and the abstract, to focus on more interesting and into the point papers. In addition, we have preferred not to use a very specific search string, dues to risk of missing papers or biasing the results. Although this has led to a rather small inclusion rate (95 out of 710), which however, despite the additional effort in IC/EC process has improved our confidence that a large pool of papers has been screen manually at the full-text level. The next step (inclusion/exclusion) has been completed very carefully, because there is always a possibility of excluding relevant articles. For avoiding this, both authors were involved in this step. Furthermore, from our searching process, we have excluded grey literature and duplicate articles and articles had been written in different language except English. The risk of bias due to missing data, based on SEGRESS [25], has been assessed as low; since: (a) we have faced no limitations with the searching space; (b) we have defined a solid process for setting the search string;

(c) we have compared the results from different DLs for identifying inconsistencies; (d) all well-known papers have been identified; (e) the study selection process was systematic; and (f) we assessed and processed all eligible papers.

Data Validity: Regarding the data validity the main threat is the subjectivity when classifying studies. This step was very time consuming so as to ensure that no mistakes are made. This step has been performed iteratively three times by the first authors, and in between iterations the classification was discussed with the second author. The same mitigation action applies to the construction of the classification schema. In our study the sample size is sufficient, and there is no publication bias since the results come from various communities. The mapping of variables to be extracted and the RQs is straightforward and have been set after a detailed discussion between the authors; the opinion of experts in the field of game design has been consulted for resolving possible terminology issues. We have not assessed the quality of the primary studies, since we have performed a SMS and not an SLR.

Research Validity: Concerning the research validity we believe that our study is reliable, because of the experience in secondary studies of the researchers and the research method is adequate for the goal of this study and no deviations from the guidelines have been performed. The results are sufficiently generalizable since they are based on a large corpus of research items. Finally, we assess the repeatability of our study as sufficient since the dataset is available, and the process is transparently described in Section 3.

7. Conclusions

There is a growing interest in game engineering, which has many differences from classic software engineering. A critical and hard to tackle issue for game developers is how they could represent the game scenario. In the literature, there are several approaches; therefore, a selection of the most suitable one is not an easy task. To alleviate this problem, in this mapping study, we: (a) categorize the scenarios' representation methods in more general categories; (b) create subcategories, when possible, based on the similar characteristics of approaches; (c) present the components and connectors of scenarios that are used in these approaches; and (d) map the part of scenario represented by each approach and component. The diagrams are the most popular way of representing scenarios, followed by algorithms, as subcategory. The objects are the most frequent components of algorithms and transitions are the most popular connectors. We believe that our findings are interesting for both researchers and practitioners in the area of computer game development. Researchers will get useful information for empirically investigating several game engineering aspects. As an example, the difficulty of implementing specific design choices may be investigated. Another example could be researching the impact of scenario design choices on the characteristics of implemented games, e.g., user satisfaction. Practitioners on the other hand, may be informed on the most common scenario design choices and combinations of design elements made before, to decide what approach to take in designing their own scenarios.

References

[1] T.M. Connolly, E.A. Boyle, E. MacArthur, T. Hainey, and J.M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers and Education*, Vol. 59, No. 2, 2012, pp. 661–686.

[2] Y.T.C. Yang, "Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation," *Computers and Education*, Vol. 59, No. 2, 2012, pp. 365–377.
[3] K.D. Squire, "Video game–based learning: An emerging paradigm for instruction," *Performance Improvement Quarterly*, Vol. 21, No. 2, 2008, pp. 7–36.
[4] H. Ham and Y. Lee, "An empirical study for quantitative evaluation of game satisfaction," in *International Conference on Hybrid Information Technology*, Vol. 2. IEEE, 2006, pp. 724–729.
[5] M.E. Paschali, A. Ampatzoglou, A. Chatzigeorgiou, and I. Stamelos, "Non-functional requirements that influence gaming experience: A survey on gamers satisfaction factors," in *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services*, 2014, pp. 208–215.
[6] M.E. Paschali, N. Bafatakis, A. Ampatzoglou, A. Chatzigeorgiou, and I. Stamelos, "Tool-assisted game scenario representation through flow charts." in *ENASE*, 2018, pp. 223–232.
[7] A. Ampatzoglou and I. Stamelos, "Software engineering research for computer games: A systematic review," *Information and Software Technology*, Vol. 52, No. 9, 2010, pp. 888–901.
[8] A.R. Teyseyre and M.R. Campo, "An overview of 3D software visualization," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 15, No. 1, 2008, pp. 87–105.
[9] F.G. Silva, "Practical methodology for the design of educational serious games," *Information*, Vol. 11, No. 1, 2019, p. 14.
[10] M. Zyda, "From visual simulation to virtual reality to games," *Computer*, Vol. 38, No. 9, 2005, pp. 25–32.
[11] P. Zemliansky and D. Wilcox, *Design and Implementation of Educational Games: Theoretical and Practical Perspectives*. IGI Global, 2010.
[12] N. Partlan, E. Carstensdottir, E. Kleinman, S. Snodgrass, C. Harteveld et al., "Evaluation of an automatically-constructed graph-based representation for interactive narrative," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, 2019, pp. 1–9.
[13] E. Segel and J. Heer, "Narrative visualization: Telling stories with data," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, 2010, pp. 1139–1148.
[14] E. Paschali, A. Ampatzoglou, R. Escourrou, A. Chatzigeorgiou, and I. Stamelos, "A metric suite for evaluating interactive scenarios in video games: an empirical validation," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 1614–1623.
[15] V. Tovinkere and M. Voss, "Flow graph designer: A tool for designing and analyzing Intel® threading building blocks flow graphs," in *43rd International Conference on Parallel Processing Workshops*. IEEE, 2014, pp. 149–158.
[16] R. Wettel and M. Lanza, "Visualizing software systems as cities," in *4th International Workshop on Visualizing Software for Understanding and Analysis*. IEEE, 2007, pp. 92–99.
[17] C. Fabricatore, "Gameplay and game mechanics: A key to quality in videogames," 2007.
[18] K. Sedig, P. Parsons, and R. Haworth, "Player–game interaction and cognitive gameplay: A taxonomic framework for the core mechanic of videogames," in *Informatics*, Vol. 4, No. 1, MDPI, 2017, p. 4.
[19] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 9, No. 1, 2013, pp. 1–22.
[20] W.Y. Chan, J. Qin, Y.P. Chui, and P.A. Heng, "A serious game for learning ultrasound-guided needle placement skills," *IEEE Transactions on Information Technology in Biomedicine*, Vol. 16, No. 6, 2012, pp. 1032–1042.
[21] J.Y. Park and J.H. Park, "A graph-based representation of game scenarios; methodology for minimizing anomalies in computer game," *The Visual Computer*, Vol. 26, No. 6, 2010, pp. 595–605.
[22] K. Kiili, "Call for learning-game design patterns," in *Educational games: Design, learning and applications*, 2010, pp. 299–311.
[23] A. Amory, "Game object model version ii: A theoretical framework for educational game development," *Educational Technology Research and Development*, Vol. 55, No. 1, 2007, pp. 51–77.

[24] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, Vol. 64, 2015, pp. 1–18.

[25] B.A. Kitchenham, L. Madeyski, and D. Budgen, "SEGRESS: Software engineering guidelines for reporting secondary studies," *IEEE Transactions on Software Engineering*, 2022.

[26] J. Guo, N. Singer, and R. Bastide, "A serious game engine for interview simulation: Application to the development of doctor-patient communication skills," in *6th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*. IEEE, 2014, pp. 1–6.

[27] P. Mildner, B. John, A. Moch, and W. Effelsberg, "Creation of custom-made serious games with user-generated learning content," in *13th Annual Workshop on Network and Systems Support for Games*. IEEE, 2014, pp. 1–6.

[28] M. Cutumisu, C. Onuczko, M. McNaughton, T. Roy, J. Schaeffer et al., "ScriptEase: A generative/adaptive programming paradigm for game scripting," *Science of Computer Programming*, Vol. 67, No. 1, 2007, pp. 32–58.

[29] H. Yin, L. Luo, W. Cai, and J. Zhong, "Data-driven dynamic adaptation framework for multi-agent training game," in *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 2. IEEE, 2015, pp. 308–311.

[30] L. Pons, C. Bernon, and P. Glize, "Scenario control for (serious) games using self-organizing multi-agent systems," in *IEEE International Conference on Complex Systems (ICCS)*. IEEE, 2012, pp. 1–6.

[31] R.C.R. Mota, D.J. Rea, A. Le Tran, J.E. Young, E. Sharlin et al., "Playing the 'trust game' with robots: Social strategies and experiences," in *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 519–524.

[32] D. Wiebusch, M. Fischbach, M.E. Latoschik, and H. Tramberend, "Evaluating scala, actors, and ontologies for intelligent realtime interactive systems," in *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, 2012, pp. 153–160.

[33] T. Schaul, "A video game description language for model-based or interactive learning," in *Conference on Computational Inteligence in Games (CIG)*. IEEE, 2013, pp. 1–8.

[34] K.A.M. Heydn, M.P. Dietrich, M. Barkowsky, G. Winterfeldt, S. von Mammen et al., "The golden bullet: A comparative study for target acquisition, pointing and shooting," in *11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. IEEE, 2019, pp. 1–8.

[35] F. Collé, R. Champagnat, and A. Prigent, "Scenario analysis based on linear logic," in *Proceedings of the ACM SIGCHI International Conference on Advances in computer entertainment technology*, 2005, pp. 1–es.

[36] T. Terzidou, T. Tsiatsos, A. Dae, O. Samaras, and A. Chasanidou, "Utilizing virtual worlds for game based learning: Grafica, a 3D educational game in second life," in *12th International Conference on Advanced Learning Technologies*. IEEE, 2012, pp. 624–628.

[37] H. Duin, M. Oliveira, and K.D. Thoben, "A methodology for developing serious gaming stories for sustainable manufacturing," in *18th International ICE Conference on Engineering, Technology and Innovation*. IEEE, 2012, pp. 1–9.

[38] U. Rüppel and K. Schatz, "Designing a BIM-based serious game for fire safety evacuation simulations," *Advanced Engineering Informatics*, Vol. 25, No. 4, 2011, pp. 600–611.

[39] J. Baldeón, I. Rodríguez, A. Puig, D. Gómez, and S. Grau, "From learning to game mechanics: The design and the analysis of a serious game for computer literacy," in *11th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2016, pp. 1–6.

[40] C. Thompson, S. Mohamed, W.Y.G. Louie, J.C. He, J. Li et al., "The robot tangy facilitating trivia games: A team-based user-study with long-term care residents," in *International Symposium on Robotics and Intelligent Sensors (IRIS)*. IEEE, 2017, pp. 173–178.

[41] G.J. Hwang, L.H. Yang, and S.Y. Wang, "A concept map-embedded educational computer game for improving students' learning performance in natural science courses," *Computers and Education*, Vol. 69, 2013, pp. 121–130.

[42] J. Wang, Z. Zhou, and M. Yu, "Pricing models in a sustainable supply chain with capacity constraint," *Journal of Cleaner Production*, Vol. 222, 2019, pp. 57–76.

[43] R.C. Nickerson, U. Varshney, and J. Muntermann, "A method for taxonomy development and its application in information systems," *European Journal of Information Systems*, Vol. 22, No. 3, 2013, pp. 336–359.

[44] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," *Information and Software Technology*, Vol. 106, 2019, pp. 201–230.

[45] X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang, "A map of threats to validity of systematic literature reviews in software engineering," in *23rd Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2016, pp. 153–160.

[46] S. Keele et al., "Guidelines for performing systematic literature reviews in software engineering," Technical report, ver. 2.3 ebse technical report. ebse, Tech. Rep., 2007.

[47] X. Xu, J. Wu, K. Fujita, T. Kato, and F. Sugaya, "Hey Peratama: A breeding game with spoken dialogue interface," in *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*, 2014, pp. 266–267.

[48] M. Gabsdil, A. Koller, and K. Striegnitz, "Natural language and inference in a computer game," in *COLING: The 19th International Conference on Computational Linguistics*, 2002.

[49] E. Ishchukova, E. Maro, and G. Veselov, "Development of information security quest based on use of information and communication technologies," in *Proceedings of the 12th International Conference on Security of Information and Networks*, 2019, pp. 1–5.

[50] T. Frtala and V. Vranic, "Animating organizational patterns," in *8th International Workshop on Cooperative and Human Aspects of Software Engineering*. IEEE, 2015, pp. 8–14.

[51] J.M. Gauthier, "Gaming: Back to the basics," in *ACM SIGGRAPH ASIA 2008 educators programme*, 2008, pp. 1–4.

[52] G. Mehlmann, B. Endrass, and E. André, "Modeling parallel state charts for multithreaded multimodal dialogues," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, 2011, pp. 385–392.

[53] A. Hautasaari, "Machine translation effects on group interaction: An intercultural collaboration experiment," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, 2010, pp. 69–78.

[54] A.L. Martin-Niedecken, K. Rogers, L. Turmo Vidal, E.D. Mekler, and E. Márquez Segura, "Exercube vs. personal trainer: Evaluating a holistic, immersive, and adaptive fitness game setup," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.

[55] T. Takahashi, K. Tanaka, and N. Oka, "Adaptive mixed-initiative dialog motivates a game player to talk with an NPC," in *Proceedings of the 6th International Conference on Human-Agent Interaction*, 2018, pp. 153–160.

[56] S. Coros, P. Beaudoin, and M. Van de Panne, "Robust task-based control policies for physics-based characters," in *ACM SIGGRAPH Asia 2009 papers*, 2009, pp. 1–9.

[57] M. Neuenhaus and M. Aly, "Empathy up," in *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 86–92.

[58] Y. Gu and M. Veloso, "Effective team-driven multi-model motion tracking," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 2006, pp. 210–217.

[59] J.F. Weng, S.S. Tseng, and T.J. Lee, "Teaching boolean logic through game rule tuning," *IEEE Transactions on Learning Technologies*, Vol. 3, No. 4, 2010, pp. 319–328.

[60] O. Janssens, K. Samyny, R. Van de Walle, and S. Van Hoecke, "Educational virtual game scenario generation for serious games," in *3rd International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 2014, pp. 1–8.

[61] J.P. David, A. Lejeune, and E. Villiot-Leclercq, "Expressing workshop scenario with computer independent model," in *Sixth International Conference on Advanced Learning Technologies*. IEEE Computer Society, 2006, pp. 1168–1169.

[62] S. Yingying, G. Liyan, and Z. Zuyao, "Researches and development of interactive educational toys for children," in *International Conference on Artificial Intelligence and Education (ICAIE)*. IEEE, 2010, pp. 344–346.

[63] A. Parakh, P. Chundi, and M. Subramaniam, "An approach towards designing problem networks in serious games," in *onference on Games (CoG)*. IEEE, 2019, pp. 1–8.
[64] G. Kontogianni and A. Georgopoulos, "A realistic gamification attempt for the Ancient Agora of Athens," in *2015 Digital Heritage*, Vol. 1. IEEE, 2015, pp. 377–380.
[65] R. Antkiewicz, W. Kulas, A. Najgebauer, D. Pierzchala, J. Rulka et al., "Selected problems of designing and using deterministic and stochastic simulators for military trainings," in *43rd Hawaii International Conference on System Sciences*. IEEE, 2010, pp. 1–10.
[66] E.L. Oliveira, D. Orru, T. Nascimento, and A. Bonarini, "Activity recognition in a physical interactive robogame," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2017, pp. 92–97.
[67] E. Ruffaldi, M. Satler, G.P.R. Papini, and C.A. Avizzano, "A flexible framework for mobile based haptic rendering," in *IEEE RO-MAN*. IEEE, 2013, pp. 732–737.
[68] I. Mayer, G. Bekebrede, C. Harteveld, H. Warmelink, Q. Zhou et al., "The research and evaluation of serious games: Toward a comprehensive methodology," *British Journal of Educational Technology*, Vol. 45, No. 3, 2014, pp. 502–527.
[69] S. Shenjie, K.P. Thomas, K.G. Smitha, and A.P. Vinod, "Two player EEG-based neurofeedback ball game for attention enhancement," in *International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 3150–3155.
[70] J. Chen, K. He, Q. Yuan, G. Xue, R. Du et al., "Batch identification game model for invalid signatures in wireless mobile networks," *IEEE Transactions on Mobile Computing*, Vol. 16, No. 6, 2016, pp. 1530–1543.
[71] H. Kharrufa, H. Al-Kashoash, and A.H. Kemp, "A game theoretic optimization of RPL for mobile Internet of Things applications," *IEEE Sensors Journal*, Vol. 18, No. 6, 2018, pp. 2520–2530.
[72] I.N. Sukajaya, A.V. Vitianingsih, S.S. Mardi, K.E. Purnama, M. Hariadi et al., "Multi--parameter dynamic difficulty game's scenario using box-muller of gaussian distribution," in *7th International Conference on Computer Science and Education (ICCSE)*. IEEE, 2012, pp. 1666–1671.
[73] H. Duin and K.D. Thoben, "Serious gaming for sustainable manufacturing: A requirements analysis," in *17th International Conference on Concurrent Enterprising*. IEEE, 2011, pp. 1–8.
[74] M. Moghim, R. Stone, P. Rotshtein, and N. Cooke, "Adaptive virtual environments: A physiological feedback HCI system concept," in *7th Computer Science and Electronic Engineering Conference (CEEC)*. IEEE, 2015, pp. 123–128.
[75] K. Schaaff and M.T. Adam, "Measuring emotional arousal for online applications: Evaluation of ultra-short term heart rate variability measures," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 362–368.
[76] K. Sekiyama, R. Carnieri, and T. Fukuda, "Strategy generation with cognitive distance in two-player games," in *International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. IEEE, 2007, pp. 166–171.
[77] Z.M. Osman, J. Dupire, S. Mader, P. Cubaud, and S. Natkin, "Monitoring player attention: A non-invasive measurement method applied to serious games," *Entertainment Computing*, Vol. 14, 2016, pp. 33–43.
[78] Y. Li, P. Su, and W. Li, "A game map complexity measure based on hamming distance," *Physics Procedia*, Vol. 22, 2011, pp. 634–640.
[79] R. Vidal and S. Sastry, "Vision-based detection of autonomous vehicles for pursuit-evasion games," *IFAC Proceedings Volumes*, Vol. 35, No. 1, 2002, pp. 391–396.
[80] G. Morgan, "Highly interactive scalable online worlds," *Advances in Computers*, Vol. 76, 2009, pp. 75–120.
[81] T. Süße and U. Wilkens, "Preparing individuals for the demands of PSS work environments through a game-based community approach–design and evaluation of a learning scenario," *Procedia CIRP*, Vol. 16, 2014, pp. 271–276.
[82] B. Sheppard, "World-championship-caliber Scrabble," *Artificial Intelligence*, Vol. 134, No. 1–2, 2002, pp. 241–275.

[83] H.Y. Sung, G.J. Hwang, and Y.F. Yen, "Development of a contextual decision-making game for improving students' learning performance in a health education course," *Computers and Education*, Vol. 82, 2015, pp. 179–190.

[84] L.F. Maia, W. Viana, and F. Trinta, "Transposition of location-based games: Using procedural content generation to deploy balanced game maps to multiple locations," *Pervasive and Mobile Computing*, Vol. 70, 2021, p. 101302.

[85] M.S. Morley, M. Khoury, and D.A. Savić, "Serious game approach to water distribution system design and rehabilitation problems," *Procedia Engineering*, Vol. 186, 2017, pp. 76–83.

[86] R. Zhao, X. Zhou, J. Han, and C. Liu, "For the sustainable performance of the carbon reduction labeling policies under an evolutionary game simulation," *Technological Forecasting and Social Change*, Vol. 112, 2016, pp. 262–274.

[87] S. Heinonen, M. Minkkinen, J. Karjalainen, and S. Inayatullah, "Testing transformative energy scenarios through causal layered analysis gaming," *Technological Forecasting and Social Change*, Vol. 124, 2017, pp. 101–113.

[88] S. O'Connor, S. Hasshu, J. Bielby, S. Colreavy-Donnelly, S. Kuhn et al., "SCIPS: A serious game using a guidance mechanic to scaffold effective training for cyber security," *Information Sciences*, Vol. 580, 2021, pp. 524–540.

[89] A.J.Q. Tan, C.C.S. Lee, P.Y. Lin, S. Cooper, L.S.T. Lau et al., "Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: A randomized controlled trial," *Nurse Education Today*, Vol. 55, 2017, pp. 38–44.

[90] T.A. Scardovelli and A.F. Frère, "The design and evaluation of a peripheral device for use with a computer game intended for children with motor disabilities," *Computer Methods and Programs in Biomedicine*, Vol. 118, No. 1, 2015, pp. 44–58.

[91] R.P. de Lope, J.R.L. Arcos, N. Medina-Medina, P. Paderewski, and F. Gutiérrez-Vela, "Design methodology for educational games based on graphical notations: Designing urano," *Entertainment Computing*, Vol. 18, 2017, pp. 1–14.

[92] Y. Pan, J. Hussain, X. Liang, and J. Ma, "A duopoly game model for pricing and green technology selection under cap-and-trade scheme," *Computers and Industrial Engineering*, Vol. 153, 2021, p. 107030.

[93] J. Radianti, M.B. Lazreg, and O.C. Granmo, "Fire simulation-based adaptation of SmartRescue App for serious game: Design, setup and user experience," *Engineering Applications of Artificial Intelligence*, Vol. 46, 2015, pp. 312–325.

[94] R.A. Agis, S. Gottifredi, and A.J. García, "An event-driven behavior trees extension to facilitate non-player multi-agent coordination in video games," *Expert Systems with Applications*, Vol. 155, 2020, p. 113457.

[95] S. Lambe, I. Knight, T. Kabir, J. West, R. Patel et al., "Developing an automated VR cognitive treatment for psychosis: gameChange VR therapy," *Journal of Behavioral and Cognitive Therapy*, Vol. 30, No. 1, 2020, pp. 33–40.

[96] H. Mitsuhara, T. Inoue, K. Yamaguchi, Y. Takechi, M. Morimoto et al., "Web-based system for designing game-based evacuation drills," *Procedia Computer Science*, Vol. 72, 2015, pp. 277–284.

[97] F. Buttussi, T. Pellis, A.C. Vidani, D. Pausler, E. Carchietti et al., "Evaluation of a 3D serious game for advanced life support retraining," *International Journal of Medical Informatics*, Vol. 82, No. 9, 2013, pp. 798–809.

[98] A. Torres, B. Kapralos, C. Da Silva, E. Peisachovich, and A. Dubrowski, "A scenario editor to create and modify virtual simulations and serious games for mental health education," in *12th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, 2021, pp. 1–4.

[99] F. Arango, C. Chang, S.K. Esche, and C. Chassapis, "A scenario for collaborative learning in virtual engineering laboratories," in *37th Annual Frontiers in Education Conference – Global Engineering: Knowledge Without Borders, Opportunities Without Passports*. IEEE, 2007, pp. F3G–7.

[100] Y. Francillette, A. Gouaich, and L. Abrouk, "Adaptive gameplay for mobile gaming," in *Conference on Computational Intelligence and Games (CIG)*. IEEE, 2017, pp. 80–87.

[101] V. Spichak and S. Petrov, "Experience in designing and developing the educational game blocksolver," in *V International Conference on Information Technologies in Engineering Education (Inforino)*. IEEE, 2020, pp. 1–5.

[102] İ. Şahin and T. Kumbasar, "Catch me if you can: A pursuit-evasion game with intelligent agents in the unity 3d game environment," in *International Conference on Electrical Engineering (ICEE)*. IEEE, 2020, pp. 1–6.

[103] M. Lohr and E. Wallinger, "Collage-the carnuntum scenario," in *Fifth IEEE International Conference on Wireless, Mobile, and Ubiquitous Technology in Education (wmute 2008)*. IEEE, 2008, pp. 161–163.

[104] Z. Ibrahim, M.C. Soo, M.T. Soo, and H. Aris, "Design and development of a serious game for the teaching of requirements elicitation and analysis," in *International Conference on Engineering, Technology and Education (TALE)*. IEEE, 2019, pp. 1–8.

[105] N.A.G. Arachchilage and M.A. Hameed, "Designing a serious game: teaching developers to embed privacy into software systems," in *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering Workshops*, 2020, pp. 7–12.

[106] B. Correia, P. Urbano, and L. Moniz, "DEVELOP-FPS: A first person shooter development tool for rule-based scripts," in *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. IEEE, 2012, pp. 1–6.

[107] Y.H. Lin, H.F. Mao, Y.C. Tsai, and J.J. Chou, "Developing a serious game for the elderly to do physical and cognitive hybrid activities," in *6th International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 2018, pp. 1–8.

[108] M. Lohr, "Mobile learning by the example of the carnuntum scenario," in *International Conference on Intelligent Networking and Collaborative Systems*. IEEE, 2009, pp. 46–52.

[109] S.H. Ab Hamid and N. Ismail, "The design of mobigp by using tamagotchi," in *First IEEE International Symposium on Information Technologies and Applications in Education*. IEEE, 2007, pp. 382–387.

[110] S. Veziridis, P. Karampelas, and I. Lekea, "Learn by playing: A serious war game simulation for teaching military ethics," in *Global Engineering Education Conference (EDUCON)*. IEEE, 2017, pp. 920–925.

[111] M.I.O. Hernández, R.M. Lezama, and S.M. Gómez, "Work-in-progress: The road to learning, using gamification." in *Global Engineering Education Conference (EDUCON)*. IEEE, 2021, pp. 1393–1397.

[112] K. Szczurowski and M. Smith, ""woodlands" – A virtual reality serious game supporting learning of practical road safety skills," in *Games, Entertainment, Media Conference (GEM)*. IEEE, 2018, pp. 1–9.

[113] L. Xu, B. Li, W. Xie, and L. Zhang, "The design and implementation of arrow game projection interactive system based on deep learning," in *International Symposium on Autonomous Systems (ISAS)*. IEEE, 2020, pp. 163–167.

[114] J.E. Almeida, J.T.P.N. Jacob, B.M. Faria, R.J. Rossetti, and A.L. Coelho, "Serious games for the elicitation of way-finding behaviours in emergency situations," in *9th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2014, pp. 1–7.

[115] F. Bellotti, R. Berta, P. Paranthaman, G. Dange, and A. De Gloria, "REAL: Reality-enhanced applied games," *IEEE Transactions on Games*, Vol. 12, No. 3, 2019, pp. 281–290.

[116] P. Herold, U. Khwaja, S. Murthy, and C. Dasgupta, "RoadEthos: Game-based learning to sensitize children on road safety through ethical reasoning," in *Tenth International Conference on Technology for Education (T4E)*. IEEE, 2019, pp. 27–33.

[117] B. Belkhouche, S. Alhadhrami, M. Alaleeli, A. Saleh, and D. Al Sharif, "Game simulation of smart taxis," in *Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 2019, pp. 1026–1031.

[118] E.P. Nunes, A.R. Luz, E.M. Lemos, C. Maciel, A.M. dos Anjos et al., "Mobile serious game proposal for environmental awareness of children," in *Frontiers in Education Conference (FIE)*. IEEE, 2016, pp. 1–8.

[119] W.M. Shalash, S. Al Tamimi, E. Abdu, and A. Barom, "No limit: A down syndrome children educational game," in *Games, Entertainment, Media Conference (GEM)*. IEEE, 2018, pp. 352–358.

[120] A. Zaraki, L. Wood, B. Robins, and K. Dautenhahn, "Development of a semi-autonomous robotic system to assist children with autism in developing visual perspective taking skills," in *27th International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 969–976.

[121] F. Grivokostopoulou, I. Perikos, and I. Hatzilygeroudis, "An educational game for teaching search algorithms," in *International Conference on Computer Supported Education*, Vol. 3. SCITEPRESS, 2016, pp. 129–136.

[122] J. Hamari, L. Keronen, and K. Alha, "Why do people play games? A review of studies on adoption and use," in *48th Hawaii International Conference on System Sciences*. IEEE, 2015, pp. 3559–3568.

[123] M.A.S. Bissaco, A.F. Frere, L.F. Bissaco, A.L. Manrique, E. Dirani et al., "A computerized tool to assess reading skills of students with motor impairment," *Medical Engineering and Physics*, Vol. 77, 2020, pp. 31–42.

Appendix A. List of primary studies

	V1	V2–V5	V6	V7	V8	V9	V10
1	A Metric Suite for Evaluating Interactive Scenarios in Video Games: An Empirical Validation	[14]	UML	flowcharts and character models	goal, game state, events	sequence, choice, flow	game story
2	Hey Peratama: A Breeding Game with Spoken Dialogue interface	[47]	narrative		goal, scene	sequence	game story
3	Natural Language and Inference in a Computer Game	[48]	logic	KL1	object, character, location (colours, accessible)	condition, navigation	game world
4	Development of Information Security Quest Based on Use of Information and Communication Technologies	[49]	narrative		location, object, goal®, character	action	game story, game world, game rules
5	Scenario Analysis Based On Linear Logic	[35]	UML, logic	Petri-Net	location, object, character	action	game story, game world
6	Animating Organizational Patterns	[50]	UML	State machine, diagrams, algorithm	scene®, character	condition, action	game rules, game story
7	Gaming: Back to the Basics	[51]	sketch		location, speed, object	navigation	game world
8	Modelling Parallel State Charts for Multithreaded Multimodal Dialogues	[52]	UML	state chart	scene, dialog, character, location, character state, animation	flow, transition	game story
9	machine translation Effects on Group Interaction: An Intercultural Collaboration Experiment	[53]	narrative		dialog	transition	game story
10	Evaluating Scala, Actors, & Ontologies for Intelligent real time interactive systems	[32]	diagram	algorithm	character, object	condition, action	game story

	V1	V2–V5	V6	V7	V8	V9	V10
11	ExerCube vs. Personal Trainer: Evaluating a Holistic, Immersive, and Adaptive Fitness Game Setup	[54]	narrative		characteristic, object, movement, characters, e®, speed®, level®, animation	transition	game story, game rules
12	Adaptive Mixed-Initiative Dialog Motivates a Game Player to Talk with an NPC	[55]	narrative		dialog, character, question		game story
13	Robust Task-based Control Policies for Physics-based Characters	[56]	pseudocode	algorithm	animation, character state	action, transition, control	game story
14	Creation of Custom-made Serious Games with User-generated Learning Content	[27]	logic	JSON schema	object, score®, level®, game state®	transition	game rules, game story
15	Empathy Up	[57]	narrative		character, score®, object	transition	game rules, game world
16	Effective Team-Driven Multi-Model Motion Tracking	[58]	UML	state machine diagrams, algorithms	object, character, move, goal, location, speed, characteristic	transition	game world, game rules, game story
17	Teaching Boolean Logic through Game Rule Tuning	[59]	logic	algorithm	event, score	action	game rules
18	Educational Virtual Game Scenario Generation for Serious Games	[60]	use case	ATTAC_L, algorithm	character, dialog®, move®, decision®, object	transition, sequence	game story, game rules
19	Playing the “Trust Game” with Robots Social Strategies and Experiences	[31]	UML	state machine diagram	dialog, move, game state	transition, control	game rules
20	Scenario control for (serious) games using self-organizing multi-agent systems	[30]	diagram	Adaptive Multi-Agent System (AMAS)	object, character state	transition	game rules
21	A Serious Game Engine for Interview Simulation: Application to the development of doctor-patient communication	[26]	use case	dialogue model	dialog, object, decision	actions	game rules

	V1	V2–V5	V6	V7	V8	V9	V10
22	Data-driven Dynamic Adaptation Framework for Multi-Agent Training Game	[29]	neural networks	algorithm	event, decision, game state, character state, goal	Adaptation Decision Feature, sequence	game rules
23	Expressing workshop Scenario with Computer Independent Model	[61]	UML	LDL	game state, character,	action	game rules
24	Researches and Development of Interactive Educational Toys for Children	[62]	narrative	algorithm	animation, object	transition	game world
25	An Approach Towards Designing Problem Networks in Serious Games	[63]	graph		goals, location, dialog, object	action, condition, navigation	game rules
26	A Realistic Gamification Attempt for the Ancient Agora of Athens	[64]	diagram	flowchart	answer®, questions®, object	action, sequence	game rules, game world
27	Selected Problems of Designing and Using Deterministic and Stochastic Simulators for Military Trainings	[65]	UML	state machine diagrams	object, event, speed, characteristic	action	game world
28	Activity recognition in a Physical Interactive	[66]	diagram	algorithms	scene	control, transition	game world
29	A flexible framework for mobile based haptic rendering	[67]	logic	algorithms	object, character state, move	action	game story
30	From Learning to Game Mechanics: The Design and the Analysis of a Serious Game for Computer Literacy	[39]	diagram	flowchart	character, level®, score®, goal, answers®, question®, time®	transition	game rules, game story
31	A Data-driven Approach for Online Adaptation of Game Difficulty	[29]	logic		event, decision®	sequence, control	game world, game rules
32	The Robot Tangy Facilitating Trivia Games: A Team based User-Study with Long-Term Care Residents	[40]	diagram	algorithms	score, time, question, answer	transition	game rules
33	Designing game methods of educational systems for maritime specialists advanced training	[68]	pseudocode	algorithms	goal®, character, object, decision®	transition	game rules, game story

	V1	V2–V5	V6	V7	V8	V9	V10
34	Two Player EEG-based Neurofeedback Ball Game for Attention Enhancement	[69]	diagram	GUI	character, object, time®	transition, control	game story, game rules
35	Batch Identification Game Model for Invalid Signatures in Wireless Mobile Networks	[70]	pseudocode, diagram	algorithms	character, dialog®, decision®	transition	game story, game rules
36	The Golden Bullet: A Comparative Study for Target Acquisition, Pointing and Shooting	[34]	narrative		object, move, animation, time	control	game story
37	A Game Theoretic Optimization of RPL for Mobile Internet of Things Applications	[71]	diagram	graph	time®, dialog®, goals®, character, animation, decision®	transition	game story, game rules
38	Multi-Parameter Dynamic Difficulty Game's Scenario Using Box-Muller of Gaussian Distribution	[72]	pseudocode	algorithm	score®, characters, level(gate)®, questions®(problem), answer®	transition	game rules, game story
39	A Methodology for Developing Serious Gaming Stories for Sustainable Manufacturing	[37]	diagram	Algorithm (LCA)	object, goal®, question®, answer®	transition, controls	game story, game rules
40	Serious Gaming for Sustainable Manufacturing: A Requirements Analysis	[73]	diagram	Algorithms (LCA)	goal®, game state®, decision®,	transition	game rules
41	Utilizing virtual worlds for game-based learning: Grafica, a 3D educational game in Second Life	[36]	diagram	algorithms	character, score®, object, location®, question®, answer®, animation(s, time®	transition	game rules, game story
42	Adaptive Virtual Environments a Physiological Feedback HCI System Concept	[74]	narrative		character state,)move	action	game story
43	Measuring Emotional Arousal for Online Applications: Evaluation of Ultra-Short Term Heart Rate Variability Measures	[75]	narrative		animation, object, score®, time®, decision®	action	game rules, game story

	V1	V2–V5	V6	V7	V8	V9	V10
44	Strategy Generation with Cognitive Distance in Two-Player Games	[76]	diagram	algorithms	score®, object, characters, character state®, time®, moves®	transition	game rules, game story
45	A concept map-embedded educational computer game for improving students' learning performance in natural science courses	[41]	diagram	map	goal, object, character, characteristic, character state®, location	action	game story, game rules, game world
46	Monitoring player attention: A non-invasive measurement method applied to serious games	[77]	narrative		object, speed®, character, score®, animation, level®, move, location	control, action	game story, game world, game rules
47	A Game Map Complexity Measure Based on Hamming Distance	[78]	logic	map	move	action	game story
48	Vision-based detection of autonomous vehicles for pursuit-evasion games	[79]	logic	algorithm	scene, move, location, speed, object	action	game story
49	Highly Interactive Scalable Online Worlds	[80]	narrative	algorithm	object, event	transition	game story
50	Designing a BIM-based serious game for fire safety evacuation simulations	[38]	diagram	SHGR	character state, animation, location, object, characters	action	game story
51	Preparing individuals for the demands of PSS work environments through a game-based community approach – design and evaluation of a learning scenario	[81]	diagram	generic model	character, object, event, time	transition	game story
52	World- championship- caliber Scrabble	[82]	narrative	MAVEN	object, score, move, location	action, control	game rules
53	Development of a contextual decision-making game for improving students' learning performance in a health education course	[83]	diagram	storyline tree	decision	action, control	game rules

	V1	V2–V5	V6	V7	V8	V9	V10
54	Transposition of Location-based Games: Using Procedural Content Generation to deploy balanced game maps to multiple locations	[84]	diagram	Weighted Graph Matching Problem	location, character, animation	action	game world
55	ScriptEase: A generative/adaptive programming paradigm for game scripting	[28]	patterns	script erase	game state, dialog, character, character state, object, event	action	game story
56	Pricing models in a sustainable supply chain with capacity constraint *	[42]	diagram	CLSA structure	character, object, decision®, characteristic	transition	game story, game rules
57	Serious Game Approach to Water Distribution System Design and Rehabilitation Problems	[85]	diagram	SeGWADE	location, object	transition	game world
58	For the sustainable performance of the carbon reduction labelling policies under an evolutionary game simulation	[86]	diagram	SD model	game state	transition	game story
59	Testing transformative energy scenarios through causal layered analysis gaming	[87]	diagram	CLA pyramid	game state	transition	game story
60	SCIPS: A serious game using a guidance mechanic to scaffold effective training for cyber security	[88]	diagram	flowchart	score, character, event, decision, goal®	transition	game rules
61	Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: A randomized controlled trial	[89]	narrative		character, object	action	game story
62	The design and evaluation of a peripheral device for use with a computer game intended for children with motor disabilities	[90]	diagram	algorithm, flowchart	character, object, move	action	game story
63	Design methodology for educational games based on graphical notations: Designing Urano	[91]	diagram	algorithm	character, scene, dialog, object, location, decision®, score®, goal	transition	game story, game world, game rules

	V1	V2–V5	V6	V7	V8	V9	V10
64	A duopoly game model for pricing and green technology selection under cap-and-trade scheme	[92]	diagram	algorithm	character, object, decision®	action	game story, game rules
65	Fire simulation-based adaptation of Smart Rescue App for serious game: Design, setup and user experience	[93]	UML		character, level, location, goal®	action	game story, game rules, game world
66	An event-driven behaviours trees extension to facilitate non-player multi-agent coordination in video games	[94]	pseudocode	tree	character, event	action, sequence	game story
67	Developing an automated VR cognitive treatment for psychosis: game Change VR therapy	[95]	diagram	algorithm	goal, character, level, location	action	game story
68	A computerized tool to assess reading skills of students with motor impairment	[9]	diagram	flowchart	character, object, goals®, location	transition	game story, game rules
69	Web-Based System for Designing Game-Based Evacuation Drills	[96]	diagram, pseudocode	state machine diagrams, algorithms	scene, goal, decision, animation	transition	game story, game world
70	Evaluation of a 3D serious game for advanced life support retraining	[97]	narrative		character, location, goal	action	game story, game world
71	A Video Game Description Language for Model-based or Interactive Learning	[33]	pseudocode	algorithms	object, level®, location, game state, goal®	transition	game rules, game story
72	A Scenario Editor to Create and Modify Virtual Simulations and Serious Games for Mental Health Education	[98]	diagram	state machine diagram	characters, animations, decision, dialog, game state	transition, choice	game story
73	A Scenario for Collaborative Learning in Virtual Engineering Laboratories	[99]	narrative		character, goal, object®, speed , level®	transition, control, choices	game rules, game story
74	Adaptive Gameplay for Mobile Gaming	[100]	UML	flowchart	object, character, location, time, question®, goal, speed	transition	game story, game rules, game world

	V1	V2–V5	V6	V7	V8	V9	V10
75	Experience in Designing and Developing the Educational Game Block Solver	[101]	diagram	flowchart	character	transition	game story
76	Catch me if you can: A pursuit-evasion game with intelligent agents in the Unity 3D game environment	[102]	logic	algorithms	character, characteristic, object, moves, location	condition	game story
77	Collage – The Carnuntum Scenario	[103]	diagram	flowchart	character, question, answer, location, score®	transition	game story, game rules
78	Design and Development of a Serious Game for the Teaching of Requirements Elicitation and Analysis	[104]	narrative		character, goal, time®, level, score®	transition	game story, game rules
79	Designing a Serious Game: Teaching Developers to Embed Privacy into Software Systems	[105]	diagram	state machine diagram	goal, character, decision	action	game story
80	DEVELOP-FPS: a First-Person Shooter Development Tool for Rule-based Scripts	[106]	diagram	state machine diagram	character, location	control, transition	game story
81	Developing a Serious Game for the Elderly to Do Physical and Cognitive Hybrid Activities	[107]	diagram		goal, answer, character, score®	control, transition	game story, game rules
82	Mobile Learning by the Example of the Carnuntum Scenario	[108]	narrative		character, location, question, answer	transition	game story
83	The Design of MobiGP by Using Tamagotchi	[109]	UML	flowchart	character, goal, time	transition	game story
84	Learn by Playing A serious war game simulation for teaching military ethics	[110]	diagram	map	character, location	transition	game story
85	Work-in-Progress: The Road to Learning, Using Gamification.	[111]	narrative		character, goal, score®, time®, level	action	game rules, game story
86	”Woodlands” – a virtual reality serious game supporting learning of practical road safety skills.	[112]	narrative		character, goal, object, location		game rules, game story, game world

	V1	V2–V5	V6	V7	V8	V9	V10
87	The Design and Implementation of Arrow Game Projection Interactive System Based on Deep Learning	[113]	logic	algorithm	object, goal, location	sequence	game story
88	Serious Games for the Elicitation of Way-finding Behaviours in Emergency Situations	[114]	narrative		goal, location, move, character, level	action, control	game story
89	REAL: Reality-Enhanced Applied Games	[115]	narrative		location, character, score®, goal	transition	game story, game rules
90	RoadEthos: Game-based learning to sensitize children on road safety through ethical reasoning	[116]	narrative		character, goal, object, move®, location	sequence, flow, transition	game story, game rules
91	Game Simulation of Smart Taxis	[117]	UML		character, object, location, move	transition	game story
92	Mobile Serious Game Proposal for Environmental Awareness of Children	[118]	narrative		character, object, goal, score®	action	game story, game rules
93	No Limit: A Down Syndrome Children Educational Game	[119]	narrative		score®, time®, level, object, move,	action	game story, game rules
94	Development of a Semi-Autonomous Robotic System to Assist Children with Autism in Developing Visual Perspective Taking Skills	[120]	diagram	state machine diagrams	event, dialog	transition	game story, game rules
95	An Educational Game for Teaching Search Algorithms	[121]	diagram	flowchart	location, move, decision, level, goal	action	game story, game rules, game world

Appendix B. Classification of studies to scenario elements

	Game story	
Time	4	[34, 81, 100, 109]
Game state	6	[14, 28, 33, 86, 87, 98]
Goal	22	[14, 39, 41, 47, 58, 91, 95–97, 99, 100, 104, 105, 107, 109, 111, 113–115]
Speed	4	[26, 79, 99, 100]
Event	5	[14, 28, 80, 81, 94]
Animation	10	[34, 36, 38, 52, 54, 56, 71, 75, 77, 98]
Move	12	[34, 67, 74, 77–79, 90, 102, 114, 117, 119, 122]
Object	36	[27, 28, 32–35, 37, 38, 41, 42, 49, 54, 58, 60, 67–69, 75–77, 79–81, 89–92, 100, 102, 112, 113, 116–119, 123]
Characteristic	5	[41, 42, 54, 58, 102]
Location	16	[33, 38, 52, 79, 95, 102, 103, 106, 108, 110, 113–117, 123]
Scene	5	[47, 52, 79, 91, 96]
Score	0	
Answer	3	[103, 107, 108]
Level	7	[93, 95, 104, 111, 114, 119, 121]
Decision	4	[96, 98, 105, 121]
Dialog	7	[28, 52, 53, 55, 91, 98, 120]
Character state	6	[28, 52, 56, 67, 74, 78]
Question	3	[55, 103, 108]
Character	52	[32, 35, 36, 38, 39, 41, 49, 50, 52, 54, 55, 58, 60, 68–72, 76, 77, 81, 84] [42, 89–95, 97–106, 123] [107–112, 114–118]
	Game rules	
Time	11	[36, 39, 40, 69, 71, 75, 76, 81, 104, 111, 119]
Game state	5	[27, 29, 31, 61, 73]
Goal	12	[29, 33, 37, 49, 63, 68, 71, 73, 88, 93, 112, 123]
Speed	2	[54, 77]
Event	4	[29, 59, 88, 120]
Animation	0	
Move	7	[26][60] [31, 76, 82, 116, 121]
Object	6	[26, 30, 36, 63, 82, 99]
Characteristic	0	
Location	3	[36, 63, 82]
Scene	1	[50]
Score	20	[27, 36, 39, 40, 57, 59, 72, 75–77, 82, 88, 91, 103, 104, 107, 111, 115, 118, 119]
Answer	6	[36, 37, 39, 40, 64, 72]
Level	7	[27, 33, 39, 54, 72, 77, 99]
Decision	13	[26, 29, 42, 60, 68, 70, 71, 73, 75, 83, 88, 91, 92]
Dialog	6	[26, 31, 60, 63, 70, 71]
Character state	4	[29, 41, 52, 76]
Question	7	[36, 37, 39, 40, 64, 72, 100]
Character	2	[61, 88]
Game World		
Time	0	
Game state	0	
Goal	0	
Speed	2	[51, 65]
Event	2	[29, 65]
Animation	3	[62, 84, 96]
Move	0	
Object	7	[48, 51, 57, 62, 64, 65, 85]

Appendix C. Conformance to SEGRESS

SEGRESS item	Discussion
Title	The paper is entitled as a mapping study
Structured abstract	Followed based on journal guidelines
Opening	1st paragraph of introduction
Rationale	2nd paragraph of introduction
Objectives	Section 3.1
Eligibility criteria	Section 3.2
Information sources	Section 3.2
Search strategy	Section 3.2
Selection process	Section 3.2
Data collection process	Section 3.3
Data items	Section 3.3
Study risk of bias assessment	Section 6
Effect measures	Not applicable
Analysis and synthesis methods	Synthesis not applicable, just classification
Reporting bias assessment	Not applicable
Certainty assessment	Not applicable
Study selection	Figure 2
Study characteristics	Section 4
Results of individual studies	Section 4
Results of analyses and synthesis	Section 4
Reporting biases	Section 6
Discussion	Section 5
Registration and protocol	The protocol is presented in Section 3

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230104, DOI: 10.37190/e-Inf230104

Story Point Estimation Using Issue Reports With Deep Attention Neural Network

Haithem Kassem*, Khaled Mahar**, Amani A. Saad***

*Multimedia Center, AASTMT, Alex, Egypt
**College of Computing and Information Technology, AASTMT, Alex, Egypt
***College of Engineering and Technology, AASTMT, Alex, Egypt

haithem_k@aast.edu, khmahar@aast.edu, amani.saad@aast.edu

Abstract

Background: Estimating the effort required for software engineering tasks is incredibly tricky, but it is critical for project planning. Issue reports are frequently used in the agile community to describe tasks, and story points are used to estimate task effort.
Aim: This paper proposes a machine learning regression model for estimating the number of story points needed to solve a task. The system can be trained from raw input data to predict outcomes without the need for manual feature engineering.
Method: Hierarchical attention networks are used in the proposed model. It has two levels of attention mechanisms implemented at word and sentence levels. The model gradually constructs a document vector by grouping significant words into sentence vectors and then merging significant sentence vectors to create document vectors. Then, the document vectors are fed into a shallow neural network to predict the story point.
Results: The experiments show that the proposed approach outperforms the state-of-the-art technique Deep-S which uses Recurrent Highway Networks. The proposed model has improved Mean Absolute Error (*MAE*) by an average of 16.6% and has improved Median Absolute Error (*MdAE*) by an average of 53%.
Conclusion: An empirical evaluation shows that the proposed approach outperforms the previous work.

Keywords: story points, deep learning, glove, hierarchical attention networks, agile, planning poker

1. Introduction

The primary goal of all software project managers is to complete the project on time and within the budget that has been established. Since the release of the agile manifesto [1], many companies have chosen to use agile approaches to guide software development. Estimating effort is critical for successful agile project management. To avoid inefficient resource allocation, accurate estimates are required [2, 3]. The story points [4, 5] are a popular method for estimating task effort. In the context of agile development, story points are typically assigned through organized group meetings known as Planning Poker sessions [6]. These meetings heavily rely on human judgment: the better the developers understand the job, the more accurate their estimates will be. Human judgment, on the other hand, is sensitive to a range of constraints. Humans are positive by nature, and

Submitted: 27 Mar. 2022; Revised: 21 Nov. 2022; Accepted: 22 Nov. 2022; Available online: 20 Jan. 2023

this bias is amplified in group interactions [7–9]. Furthermore, the presence of a project manager, other senior developers, or dominant personalities in the meeting has been shown to affect developer estimation [10].

The use of machine learning regessors has three advantages. To begin with, the regressors have a thorough understanding of the project that dates back to its beginnings, and it based its predictions on all past issues in the issue tracking system. Second, because the regressors' estimations can be tracked back to the regressor's characteristics, it is not influenced or coerced by others. Third, the estimation is repeatable and predictable: the system never grows bored of producing the same results over and over again.

We introduce a prediction model that helps teams by providing a story-point estimate for a certain user story. The model uses the team's previous story point assessments to forecast the complexity of new issues. The team's existing estimation techniques will be used in conjunction with (rather than in place of) this prediction system. It could also be used as a decision-making tool and help with estimating. This is similar to the notion of combination-based effort estimating [11, 12]. Estimates are generated from various sources, such as a combination of expert and formal model-based estimates.

The suggested model automatically learns semantic features that represent the meaning of user stories or issue reports, removing the need for users to develop and extract features manually. Feature engineering is often done by domain specialists who use their in-depth understanding of the data to develop features that machine learning algorithms may exploit. Our model is a full end-to-end system that estimates story points by passing raw data signals (i.e., words) from input nodes to the final output node. The use of hierarchical attention networks (HAN) for story point prediction is a fundamental innovation in our method.

An empirical evaluation was conducted to answer the following research questions:

RQ1. Does the use of Hierarchical Attention Networks provide more accurate story point estimates than Recurrent Highway Nets?

RQ2. Does the use of BERT provide more accurate story point estimates than using HAN?

The remainder of the paper is organized as follows: Section 2 provides context for Story Points, Planning Poker, Deep learning, and Hierarchical Attention Network. Section 3 presents related works, while Section 4 focuses on the design of the proposed model. Section 5 discusses the proposed model evaluation , Section 6 comparing with the state of art, Section 7 shows future work, and finally Section 8 presents the conclusion.

2. Background

2.1. Story points

Story points are a unit of measure for expressing the overall size of a user story,feature or another piece of work [4]. The number of story points is an indication of how difficult a specific task is for the development team, rather than measuring the quantity of work required to achieve it. As a first stage, the team normally decides on the number of story points that a baseline activity deserves. After that, estimating effort is dependent on comparison to that baseline. The Fibonacci sequence (i.e., 1, 2, 3, 5, 8, 13, 21, 34, 55, …) is

commonly deviated from when assigning story points. The uncertainty that comes with estimating complex tasks in real-world software is shown in this series.

2.2. Planning pokers

The majority of software projects rely completely on human judgment to estimate effort [13]. The most prevalent effort estimation approaches based on human judgment are those based on group estimation. When it comes to estimating story points, Planning Poker [13] is the most commonly used method. To perform Planning Poker, the customer must first communicate an issue that they would like to get handled. The developers then gather for a poker game in which each player selects a card with the desired story points for each issue to be estimated, and then all the cards are revealed at the same time. The developer that provides the lowest and highest estimate must justify their choice, thus eventually triggering further discussion which is followed by another group estimation. The process continues until the team agrees upon a consensus estimate.

2.3. Deep learning

Deep learning technology (DL) has shown impressive results in a variety of fields, including machine vision [14], speech recognition [15], and text classification [16]. Researchers can divide deep learning research on text classification into two steps: The first step is to learn word vector representations through neural language models [17], and the second step is to perform classification composition over the learned word vectors.

Deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [18] are commonly used in text classification. Recently, several text classification methods based on CNNs or RNNs have been proposed [19, 20]. The CNN learns local responses using temporal or spatial data, but cannot learn sequential correlations. RNNs, in contrast, are designed to do sequential modeling, but cannot extract features in a parallel manner.

2.4. Hierarchical attention network

Yang et al. [21] developed Hierarchical Attention Network (HANs), a DL document classification model based on RNNs. They are made up of hierarchies, with the lower hierarchies' outputs becoming the upper hierarchies' inputs. This is based on the intuition that documents are made up of meaningful sentences, which are made up of meaningful word sequences. Each HAN hierarchy is made up of a bidirectional dynamic Long short-term memory (LSTM) or gated recurrent unit (GRU) with attention mechanisms. When processing words/sentences, directionality is required so that the network can account for the prior and subsequent context. The attention mechanism is added to enable the network to put extra focus on the LSTM/GRU outputs associated with the words and lines that are most indicative of a particular class. LSTMs/GRUs are used because they allow the network to selectively process input information based on how relevant it is to the classification tasks; similarly, the attention mechanism is added to enable the network to put extra focus on the LSTM/GRU outputs associated with the words and lines that are most indicative of a particular class. Hierarchical attention networks were used to

learn semantic features that automatically convey the true meaning of user stories and predict the estimated story point.We'll go over the details of each component later in this paper.

3. Related work

Methods for estimating software work can be divided into three categories: expert-based, model-based, and hybrid techniques. Expert-based methods, which rely on human understanding to create estimates, are the most widely used technique [22, 23]. Expert-based estimation necessitates the presence of experts at all times when an estimate is required. Model-based approaches draw on data from previous projects, but they differ in terms of how they construct customized models. A fixed model in which elements and variables are fixed is the well-known construction cost (CO-COMO) model [24]. Their relationship has already been established. These estimation models were built using data from a range of past studies. As a result, they are usually only effective for the type of project that was used to develop the model. Regression (e.g., [25, 26]), neural networks (e.g., [27, 28]), fuzzy logic (e.g., [29]), Bayesian belief networks (e.g., [30]), analogy-based (e.g., [31, 32]), and multi-objective evolutionary approaches are all used in the customized model construction process (e.g., [33]). However, no single strategy is expected to perform well across all project types [34–36]. As a result, some recent research [37] suggests integrating estimates from several estimators. Papers [38, 39], which are similar to the ideas in this paper, hybrid techniques integrate expert judgments with available data.

While the majority of existing research focuses on estimating a project as a whole, less attention is paid to developing models for agile projects in particular. Different planning and estimation methodologies are required for today's agile, dynamic, and incremental projects [40]. Machine learning techniques are being used in recent approaches to assist in estimating effort for agile projects. The study recently provided an approach for extracting TF-IDF features from the problem description to construct a model for story point estimations, which was published in [41]. The retrieved features are then subjected to the uniform selection process and input into regressors such as SVM.

In addition, Cosmic Function Points (CFP) [42] estimate the effort required to finish an agile project [43]. Abrahamsson [44] created a regression model and neural network-based effort prediction model for the creation of iterative software. Unlike standard effort estimate models, this model is developed after each iteration (rather than at the end of the design phase) to estimate the effort for the next iteration.

The authors of [45] developed a Bayesian network model for estimating effort in agile Extreme Programming software projects. Their model, on the other hand, is based on several criteria (such as process effectiveness and improvement) that necessitate a significant amount of learning and fine-tuning. Bayesian networks are frequently used in [46] to model dependencies between multiple aspects in Scrum-based software development projects to identify difficulties (e.g., sprint progress and sprint planning quality affect product quality).

Choetkiertikul [47] focuses on estimating issues with story points, which is a substantial improvement over earlier work, by applying deep learning techniques to automatically learn semantic features that reflect the underlying meaning of issue descriptions. The previous study has been done in projecting the elapsed time for correcting a bug or the danger of addressing an issue with a pause (see [48–51]).

The proposed model uses pre-trained embedding vectors and transfer learning with GloVe to save training time, which is the key difference from [47]. Word-to-vector (Word2Vec) and global vector (GloVe) are two recent techniques that are well recognized for producing vector representations [52, 53]. Pennington et al. [54]. demonstrated that GloVe outperforms Word2Vec since Word2Vec has a low vector dimensionality and cannot incorporate all of the corpus data. The GloVe, in comparison, has both local and worldwide information about the words that have appeared. GloVe algorithm uses the statistics of word-word co-occurrences in a corpus and is used for similarity and entity identification [55].

Choetkiertikul [47] is the first model providing end-to-end trainable from raw input data to prediction outcomes without any manual feature engineering and has outperformed previous work [41, 43–46]. The proposed model is aiming to use deep learning and make use of the hierarchical attention mechanism, this model has the ability to detect important words and sentences. The Hierarchical Attention Network (HAN) was implemented with the goal of capturing two fundamental ideas about document organization. First, because documents are hierarchical (words make sentences, sentences form a document), we generate a document representation by first creating sentence representations and then aggregating them into a document representation. Second, different words and sentences in a document are found to have varying levels of information. The model constructs a document vector progressively by aggregating important words into sentence vectors and then aggregating important sentence vectors to document vectors.

4. The proposed model

The general goal of our research is to create a prediction system that takes the title and description of an issue as input and generates the estimated story-point. The proposed model introduces the use of hierarchical attention networks (HAN). An embedding layer, attention layers, and encoders are all components of the HAN model, which together help the model understand the textual features. The extraction of relevant context is the responsibility of the encoders. The attention layers evaluate how important a sequence of tokens is with reference to the document. The HAN essentially consists of "hierarchies," where the outputs of the lower hierarchies serve as the inputs for the upper hierarchies. We first break down a document into sentences before feeding it into the HAN. Each sentence is encoded into a vector representation using a word encoder (a bidirectional GRU) and a word attention mechanism. These sentence representations are passed through a sentence encoder with a sentence attention mechanism resulting in a document vector representation. A fully connected layer with the appropriate activation function receives this final representation and uses it to make predictions. The term "hierarchical" refers to a document's "semantic hierarchy." The same algorithms are used twice, once at the word level and once at the sentence level. The model gradually builds a document vector by grouping significant words into sentence vectors and then merging important sentence vectors to create document vectors. The document vectors are then fed into a shallow neural network to predict the story point. The proposed model is made up of five layers, as shown in Figure 1, and is explained briefly as follows:

1. Input layer: Accepts a document that is made up of sentences, each of which is made up of a series of word IDs that represent user stories or issues that describe what has to be produced in the software project. Assume that a document includes L sentences s_i and each sentence contains T_i words. w_{i_t} with $t \in [1, T]$ represents the words in the i^{th} sentence.

2. Embedding layer: Each word in each sentence is individually embedded, resulting in Sequences of word vectors, one for each sentence. It does this by converting input text into dense word vectors that encode both the meaning and context of the text. Word Representation using Global Vectors Each word's vector representation was obtained using GloVe [56]. GloVe is an unsupervised learning technique for obtaining word vector representations.

3. Encoding layer: We have a sequence of word vectors from the previous layer, and this layer seeks to compute a sentence matrix from which we can construct a document matrix. The sentence matrix is made up of rows, each representing the meaning of a single token in the phrase. A Bidirectional RNN is used to implement this layer. The vector of each token is divided into two portions, one computed with a forward pass and the other with a backward pass. To get the entire vector, we just join the two. There are two encoders in this layer:

Sentence Encoder: Converts sequence of word vectors to sentence matrix.

Document Encoder: Converts sequence of sentence vectors to document matrix.

Given a sentence with words $w_{it}, t \in [0, T]$, we first embed the words to vectors through an embedding matrix $w_e, x_{ij} = w_e w_{ij}$.

We obtain word annotations using bidirectional GRU by combining input from both directions and adding contextual information to the annotation.. The bidirectional GRU contains the forward GRU $\overrightarrow{f}$ which reads the sentence s_i from w_{i1} to w_{iT} and a backward GRU $\overleftarrow{f}$ which reads from w_{iT} to w_{i1}.

$$X_{it} = W_e w_{it}, t \in [1, T]. \tag{1}$$

$$\overrightarrow{h_i t} = \overrightarrow{GRU} X_{it}, t \in [1, T]. \tag{2}$$

$$\overleftarrow{h_i t} = \overleftarrow{GRU} X_{it}, t \in [T, 1]. \tag{3}$$

Document Encoder in a similar manner, given the sentence vectors s_i, we can obtain a document vector. To encode the sentences, we use a bidirectional GRU:

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i), i \in [1, L]. \tag{4}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i), i \in [L, 1]. \tag{5}$$

4. Attention layer: Our goal in this layer is to reduce the Sentence matrix from the previous layer to a single vector that the feed-forward network may use for prediction. This layer's job is to determine the words that are most important to a user story's meaning. The following equations were utilized in this layer [57] .

$$e_t = \tanh(Uc + Wh_t + b) \tag{6}$$

$$\alpha_t = \text{softmax}(e_t) \tag{7}$$

$$o = \sum \alpha_t h_t \tag{8}$$

c is the vector obtained by applying max poling to the matrix obtained from GRU. U is a new weight.

Output layer: We use a feedforward neural network with a linear activation function as the final regressor to construct a story-point estimate. The following is a definition for

this function:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i \tag{9}$$

where y is the output story point, x_i is an input signal from the previous layer, a_i is the trained coefficient (weight), and n is the size of the embedding dimension.

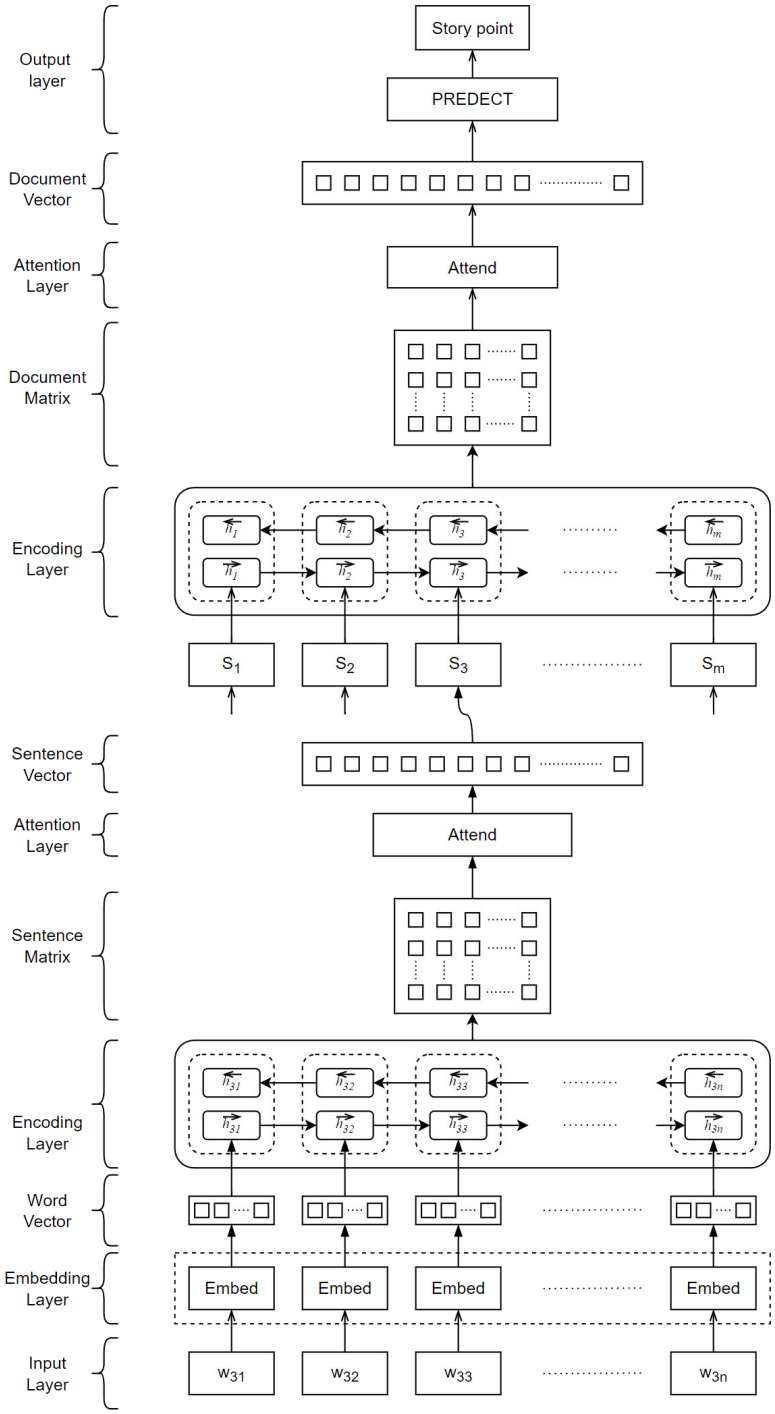

Figure 1. The proposed model

5. The proposed approach evaluation

Our data set [47] includes 23,313 issues from 16 different projects, including Apache Mesos (ME), Apache Usergrid (UG), Appcelerator Studio (AS), Aptana Studio (AP), Titanium SDK/CLI (TI), DuraCloud (DC), Bamboo (BB), Clover (CV), JIRA Software (JI), Moodle (MD), Data Management (DM), Mule (MU), Mule Studio (MS), Spring XD (XD), Talend Data Quality (TE) as shown in Table 1. The data set is divided into three parts: 60% for training, 20% for validation, and the remaining 20% for testing. (Our dataset & code are available at https://doi.org/10.5281/zenodo.7341235). In planning poker, the story points are typically ordered in a Fibonacci sequence, such as 1, 2, 3, 5, 8, 13, 21, and so on. To divide documents into sentences and tokenize each sentence, we used Natural Language Toolkit (NLTK) [58]. The suggested technique employs a Transfer Learning model called pre-trained word embedding. The basic concept is to leverage publicly available embeddings that have been trained on large datasets. Instead of randomly initializing our neural network weights, we use these previously trained integrations as initialization weights. This speeds up training and improves the performance of NLP models. The most widely used method for obtaining word embedding from text corpora is GloVe [57]. It offers pre-trained embedding based on massive text corpora. GloVe allows for different sizes of embedding. The experiment was conducted with a fifty-embedding size. When evaluating the accuracy of an effort estimating model, a variety of metrics are used. The majority of them (i.e., $|ActualSP - EstimatedSP|$) are based on the Absolute Error, where $AcutalSP$ denotes the actual story points awarded to a problem, and $EstimatedSP$ denotes the result of estimation. Prediction at level [59], i.e., Pred(l), and Mean of Magnitude of Relative Error (MRE) or Mean Percentage Error have also been employed in effort estimate. However, several investigations [59–62] have discovered that these metrics have a proclivity for underestimation and are not reliable. Consequently, the Mean Absolute Error (MAE) and Median Absolute Error ($MdAE$) have been recommended to compare

Table 1. Descriptive statistics of story point dataset

Repo.	Project	Abb.	# Issues	Min SP	Max SP	Mean SP	Median SP	Mode SP	Var SP	Std SP	Mean TD length	LOC
Apache	Mesos	ME	1680	1	40	3.09	3	3	5.87	2.42	181.12	247,542+
	Usergrid	UG	482	1	8	2.85	3	3	1.97	1.40	108.60	639,110+
Appcelerator	Appcelerator Studio	AS	2919	1	40	5.64	5	5	11.07	3.33	124.61	2,941,856#
	Aptana Studio	AP	829	1	40	8.02	8	8	35.46	5.95	124.61	6,536,521+
	Titanium SDK/CLI	TI	2251	1	34	6.32	5	5	25.97	5.10	205.90	882,986+
DuraSpace	DuraCloud	DC	666	1	16	2.13	1	1	4.12	2.03	70.91	88,978+
Atlassian	Bamboo	BB	521	1	20	2.42	2	1	4.60	2.14	133.28	6,230,465#
	Clover	CV	384	1	40	4.59	2	1	42.95	6.55	124.48	890,020#
	JIRA Software	JI	352	1	20	4.43	3	5	12.35	3.51	114.57	7,070,022#
Moodle	Moodle	MD	1166	1	100	15.54	8	5	468.53	21.65	88.86	2,976,645+
Lsstcorp	Data Management	DM	4667	1	100	9.57	4	1	275.71	16.61	69.41	125,651*
Mulesoft	Mule	MU	889	1	21	5.08	5	5	12.24	3.50	81.16	589,212+
	Mule Studio	MS	732	1	34	6.40	5	5	29.01	5.39	70.99	16,140,452#
Spring	Spring XD	XD	3526	1	40	3.70	3	1	10.42	3.23	78.47	107,916+
Talendforge	Talend Data Quality	TD	1381	1	40	5.92	5	8	26.96	5.19	104.86	1,753,463#
	Talend ESB	TE	868	1	13	2.16	2	1	2.24	1.50	128.97	18,571,052#
	Total		23,313									

effort estimation performance [63, 64] models. The term MAE is defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{n} ActualSP_i - EstimatedSP_i \tag{10}$$

where N is the number of issues used for evaluating the performance (i.e., test set), $ActualSP_i$ is the actual story point, and $EstimatedSP_i$ is the estimated story point, for the issue i. We also report the Median Absolute Error since it is more robust to large outliers. $MdAE$ is defined as

$$MdAE = \text{Median}\{|ActualSP_i - EstimatedSP_i|\} \tag{11}$$

where $1 \leq i \leq N$.

5.1. Results analysis and discussion

To compare the proposed regressor with the state of the art, we can refer to the work by Choetkiertikul [47]. They use deep learning approaches to automatically learn semantic characteristics that reflect the underlying meaning of issue descriptions to estimate issues using story points, which is a significant advance over previous work. To reduce the risk of external validity, we examined 23,313 issues across sixteen open source projects, each with its size as shown in Figure 2, complexity, development team, and community. Table 2 shows MAE and $MdAE$, achieved from hierarchical attention networks (HAN) against Deep-SE using Recurrent Highway Networks for deep representation of issue reports [47], the proposed model improved MAE between 0.7 to 28 percent over Deep-SE and Improved $MdAE$ between 18 to 68 percent over Deep-SE . Regardless of the size of the data, the improvement is noticeable. The proposed approach surpasses the previous best baseline methods by 16.5 percent and 19.4 percent for small projects like Apache Usergrid and Clover, respectively. This observation holds true across a variety of larger projects. As seen in Table 1, HAN is the best technique, continuously outperforming Deep-SE across all sixteen projects. RQ1 is answered by this finding.

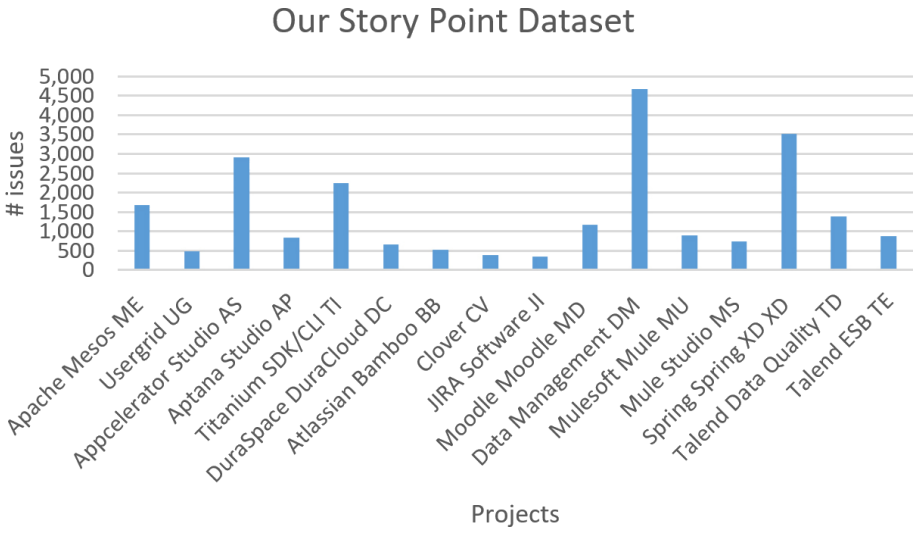

Figure 2. Story point dataset

Table 2. Comparison between the proposed model and Deep-SE

Project	Method	*MAE*	*MdAE*
Apache Mesos	Deep-SE	1.02	0.73
	HAN	**0.93**	**0.39**
Apache Usergrid	Deep-SE	1.03	0.80
	HAN	**0.84**	**0.47**
Appcelerator Studio	Deep-SE	1.36	0.56
	HAN	**1.35**	**0.54**
Aptana Studio	Deep-SE	2.71	2.52
	HAN	**2.63**	**1.13**
Titanum	Deep-SE	1.97	1.34
	HAN	**1.70**	**0.54**
DuraCloud	Deep-SE	0.68	0.53
	HAN	**0.6**	**0.12**
Bamboo	Deep-SE	0.74	0.61
	HAN	**0.67**	**0.22**
JIRA Software	Deep-SE	1.38	1.09
	HAN	**1.27**	**0.43**
Moodle	Deep-SE	5.97	4.93
	HAN	**5.66**	**1.56**
Data Management	Deep-SE	3.77	2.22
	HAN	**3.63**	**1.03**
Mule	Deep-SE	2.18	1.96
	HAN	**1.86**	**0.81**
Mule Studio	Deep-SE	3.23	1.99
	HAN	**2.56**	**1.39**
Spring XD	Deep-SE	1.63	1.31
	HAN	**1.20**	**0.51**
Talend Data Quality	Deep-SE	2.97	2.92
	HAN	**2.49**	**1.14**
Talend	Deep-SE	0.64	0.59
	HAN	**0.60**	**0.25**
Clover	Deep-SE	2.11	0.8
	HAN	**1.81**	**0.54**

Table 3. Comparison between the proposed model and BERT

Project	Method	*MAE*
Apache Mesos	HAN	0.93
	BERT	3.39
Apache Usergrid	HAN	0.84
	BERT	3.24
Appcelerator Studio	HAN	1.35
	BERT	2.50
Aptana Studio	HAN	2.63
	BERT	4.18
Titanum	HAN	1.7
	BERT	3.49
DuraCloud	HAN	0.49
	BERT	3.79
Bamboo	HAN	0.67
	BERT	2.76
JIRA Software	HAN	1.27
	BERT	3.13
Moodle	HAN	5.66
	BERT	11.99
Data Management	HAN	3.63
	BERT	7.78
Mule	HAN	1.86
	BERT	3.51
Mule Studio	HAN	2.56
	BERT	3.51
Spring XD	HAN	1.2
	BERT	3.16
Talend Data Quality	HAN	2.49
	BERT	4.04
Talend	HAN	0.60
	BERT	3.42
Clover	HAN	1.81
	BERT	3.87

To compare the performance of two estimating models, we used the Wilcoxon Signed Rank Test [65] to determine the statistical significance of the mean absolute errors obtained by the two models. Because it makes no assumptions about underlying data distributions, the Wilcoxon test is a robustness test. In order to evaluate if there were a good effect of the proposed model for estimating the effort needed for a story point, Wilcoxon signed-rank tests revealed a statistically positive change in effort estimation, $z = -3.517, p = 0.001$ with a medium effect size ($d = 0.6$), Cohen's d effect sizes [66] were calculated. Effect sizes of 0.2 were regarded as small, 0.5 as a medium, and 0.8 as large. So the HAN has medium effect size on *MAE*.

5.2. Threats to validity

We attempted to reduce the validity challenges by using real-world data from issues reported in large open-source projects. These issue reports' titles and descriptions and the actual story points assigned to them were gathered. We are aware that those story points were calculated by human teams, which means they may contain biases and, in some cases, be inaccurate. Datasets of various sizes were used in our study. Additionally, in order to reduce conclusion instability we carefully adhered to current best practices when evaluating effort estimation models. To mitigate threats to external validity, we examined 23,313 issues from sixteen open source projects that differ greatly in size, complexity, developer team, and community. We acknowledge, however, that our dataset would not be representative of all types of software projects, particularly in commercial settings (despite the fact that open-source and commercial projects are similar in many ways). The nature of contributors, developers, and project stakeholders is one of the key differences between open-source and commercial projects that may influence story point estimation. More research is required for commercial agile projects.

6. Comparing with the state of art

The Bidirectional Encoder Representations from Transformers (BERT) is a novel approach and is regarded as the cutting edge of pre-trained language representation [67]. BERT models are regarded as contextualized or dynamic models, and they have produced noticeably better results in a number of NLP tasks [68–70], including sentiment classification, calculating the semantic similarity of texts, and identifying tasks involving textual linking. The authors of [52] proposed the use of Bert for effort estimation and their experiments were conducted on the same data set. When comparing our experimental results to [52], the experimental results presented in Table 3 showed that HAN models achieved significantly higher results than the BERT model. RQ2 is answered by this finding. We conclude that a model's performance is dependent on the task and the data, so these factors should be considered before choosing a model rather than just going with the most widely used model.

7. Future work

Future work will involve comparing the results of our model to other pre-trained language models such as GPT [70] and XLNet [71]. These models have been shown to be state of the art in a variety of tasks such as question answering, named entity recognition, and natural language inference. It is also planned to test the proposed model on other Agile data sets.

8. Conclusion

The key novelty of the proposed model is using pre-trained embedding vectors and transfer learning with GloVe to reduce training time instead of creating a new embedding vector. Introducing the use of The Hierarchical Attention Network. The Hierarchical Attention Network (HAN) was created to capture two key concepts in document organization. To begin, we construct a document representation by first creating sentence representations and then

aggregating them into a document representation because documents are hierarchical (words make sentences, sentences make a document). Second, different degrees of information are discovered in different words and phrases in a document. The approach builds a document vector by first aggregating key words into sentence vectors, then aggregating important sentence vectors into document vectors. This process has a significant effect on story point prediction. The results of our experiments show that the proposed model improved *MAE* by 0.7 to 28 percent compared to Deep learning model for Story Point Estimation (Deep-SE) and *MdAE* by 18 to 68 percent compared to Deep-SE. The proposed approach regularly outperforms earlier work. The model can better locate and extract critical information.

References

[1] K. Beck, M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham et al., *Manifesto for agile software development*, 2001.

[2] L.C. Briand, "On the many ways software engineering can benefit from knowledge engineering," in *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering*, 2002, pp. 3–6.

[3] J.W. Paulson, G. Succi, and A. Eberlein, "An empirical study of open-source and closed-source software products," *IEEE Transactions on Software Engineering*, Vol. 30, No. 4, 2004, pp. 246–256.

[4] M. Cohn, "Agile estimating and planning Pearson education," 2006.

[5] S. Porru, A. Murgia, S. Demeyer, M. Marchesi, and R. Tonelli, "Estimating story points from issue reports," in *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2016, pp. 1–10.

[6] J. Grenning, "Planning poker or how to avoid analysis paralysis while release planning," *Hawthorn Woods: Renaissance Software Consulting*, Vol. 3, 2002, pp. 22–23.

[7] R. Brown and S. Pehrson, *Group processes: Dynamics within and between groups*. John Wiley & Sons, 2019.

[8] A.R. Lindesmith, A. Strauss, and N.K. Denzin, *Social psychology*. Sage, 1999.

[9] S. Nolen-Hoeksema, B. Fredrickson, G.R. Loftus, and C. Lutz, *Introduction to psychology*. Cengage Learning Washington, 2014.

[10] J. Aranda and S. Easterbrook, "Anchoring and adjustment in software estimation," in *Proceedings of the 10th European Software Engineering Conference Held Jointly With 13th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 2005, pp. 346–355.

[11] M. Jorgensen and M. Shepperd, "A systematic review of software development cost estimation studies," *IEEE Transactions on Software Engineering*, Vol. 33, No. 1, 2006, pp. 33–53.

[12] K. Moharreri, A.V. Sapre, J. Ramanathan, and R. Ramnath, "Cost-effective supervised learning models for software effort estimation in agile environments," in *40th Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2. IEEE, 2016, pp. 135–140.

[13] K. Moløkken-Østvold, N.C. Haugen, and H.C. Benestad, "Using planning poker for combining expert estimates in software projects," *Journal of Systems and Software*, Vol. 81, No. 12, 2008, pp. 2106–2117.

[14] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image and Vision Computing*, Vol. 65, 2017, pp. 15–22.

[15] K. Marasek et al., "Deep belief neural networks and bidirectional long-short term memory hybrid for speech recognition," *Archives of Acoustics*, Vol. 40, No. 2, 2015, pp. 191–195.

[16] K.S. Tai, R. Socher, and C.D. Manning, "Improved semantic representations from tree--structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[17] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, Vol. 61, 2015, pp. 85–117.

[18] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, Vol. 25, 2012.

[19] K.I. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks," *Neural Networks*, Vol. 6, No. 6, 1993, pp. 801–806.
[20] Y. Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015.
[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola et al., "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.
[22] M. Jørgensen, "A review of studies on expert estimation of software development effort," *Journal of Systems and Software*, Vol. 70, No. 1–2, 2004, pp. 37–60.
[23] M. Jørgensen and T.M. Gruschke, "The impact of lessons-learned sessions on effort estimation and uncertainty assessments," *IEEE Transactions on Software Engineering*, Vol. 35, No. 3, 2009, pp. 368–383.
[24] B. Boehm, *Software cost estimation with COCOMO II*. New Jersey: Prentice-Hall, 2000.
[25] P. Sentas, L. Angelis, and I. Stamelos, "Multinomial logistic regression applied on software productivity prediction," in *9th Panhellenic Conference in Informatics*, 2003, pp. 1–12.
[26] P. Sentas, L. Angelis, I. Stamelos, and G. Bleris, "Software productivity and effort prediction with ordinal regression," *Information and Software Technology*, Vol. 47, No. 1, 2005, pp. 17–29.
[27] S. Kanmani, J. Kathiravan, S.S. Kumar, and M. Shanmugam, "Neural network based effort estimation using class points for OO systems," in *International Conference on Computing: Theory and Applications (ICCTA'07)*. IEEE, 2007, pp. 261–266.
[28] A. Panda, S.M. Satapathy, and S.K. Rath, "Empirical validation of neural network models for agile software effort estimation based on story points," *Procedia Computer Science*, Vol. 57, 2015, pp. 772–781.
[29] S. Kanmani, J. Kathiravan, S.S. Kumar, and M. Shanmugam, "Class point based effort estimation of oo systems using fuzzy subtractive clustering and artificial neural networks," in *Proceedings of the 1st India Software Engineering Conference*, 2008, pp. 141–142.
[30] S. Bibi, I. Stamelos, and L. Angelis, "Software cost prediction with predefined interval estimates," in *Proceedings of Software Measurement European Forum*, Vol. 4, 2004, pp. 237–246.
[31] M. Shepperd and C. Schofield, "Estimating software project effort using analogies," *IEEE Transactions on Software Engineering*, Vol. 23, No. 11, 1997, pp. 736–743.
[32] L. Angelis and I. Stamelos, "A simulation tool for efficient analogy based cost estimation," *Empirical Software Engineering*, Vol. 5, No. 1, 2000, pp. 35–68.
[33] F. Sarro, A. Petrozziello, and M. Harman, "Multi-objective software effort estimation," in *38th International Conference on Software Engineering (ICSE)*. IEEE, 2016, pp. 619–630.
[34] M. Jorgensen and M. Shepperd, "A systematic review of software development cost estimation studies," *IEEE Transactions on Software Engineering*, Vol. 33, No. 1, 2006, pp. 33–53.
[35] E. Kocaguneli, T. Menzies, and J.W. Keung, "On the value of ensemble effort estimation," *IEEE Transactions on Software Engineering*, Vol. 38, No. 6, 2011, pp. 1403–1416.
[36] F. Collopy, "Difficulty and complexity as factors in software effort estimation," *International Journal of Forecasting*, Vol. 23, No. 3, 2007, pp. 469–471.
[37] E. Kocaguneli, T. Menzies, and J.W. Keung, "On the value of ensemble effort estimation," *IEEE Transactions on Software Engineering*, Vol. 6, No. 38, 2012, pp. 1403–1416.
[38] R. Valerdi, "Convergence of expert opinion via the wideband Delphi method," in *21st Annual International Symposium of the International Council on Systems Engineering, INCOSE*, Vol. 2011, 2011.
[39] S. Chulani, B. Boehm, and B. Steece, "Bayesian analysis of empirical software engineering cost models," *IEEE Transactions on Software Engineering*, Vol. 25, No. 4, 1999, pp. 573–583.
[40] M. Cohn, *Agile estimating and planning*. Pearson Education, 2005.
[41] S. Porru, A. Murgia, S. Demeyer, M. Marchesi, and R. Tonelli, "Estimating story points from issue reports," in *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2016, pp. 1–10.
[42] C. Commeyne, A. Abran, and R. Djouab, "Effort estimation with story points and cosmic function points – An industry case study," *Software Measurement News*, Vol. 21, No. 1, 2016, pp. 25–36.

[43] G. Poels, "Definition and validation of a COSMIC-FFP functional size measure for object-oriented systems," in *Proc. 7th Int. ECOOP Workshop Quantitative Approaches OO Software Eng. Darmstadt*, 2003.
[44] P. Abrahamsson, R. Moser, W. Pedrycz, A. Sillitti, and G. Succi, "Effort prediction in iterative software development processes – Incremental versus global prediction models," in *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*. IEEE, 2007, pp. 344–353.
[45] P. Hearty, N. Fenton, D. Marquez, and M. Neil, "Predicting project velocity in XP using a learning dynamic bayesian network model," *IEEE Transactions on Software Engineering*, Vol. 35, No. 1, 2008, pp. 124–137.
[46] M. Perkusich, H.O. De Almeida, and A. Perkusich, "A model to detect problems on scrum-based software development projects," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 1037–1042.
[47] M. Choetkiertikul, H.K. Dam, T. Tran, T. Pham, A. Ghose et al., "A deep learning model for estimating story points," *IEEE Transactions on Software Engineering*, Vol. 45, No. 7, 2018, pp. 637–656.
[48] E. Giger, M. Pinzger, and H. Gall, "Predicting the fix time of bugs," in *Proceedings of the 2nd International Workshop on Recommendation Systems for Software Engineering*, 2010, pp. 52–56.
[49] L.D. Panjer, "Predicting eclipse bug lifetimes," in *Fourth International Workshop on Mining Software Repositories (MSR'07: ICSE Workshops 2007)*. IEEE, 2007, pp. 29–29.
[50] P. Bhattacharya and I. Neamtiu, "Bug-fix time prediction models: Can we do better?" in *Proceedings of the 8th Working Conference on Mining Software Repositories*, 2011, pp. 207–210.
[51] P. Hooimeijer and W. Weimer, "Modeling bug report quality," in *Proceedings of the Twenty-Second IEEE/ACM International Conference on Automated Software Engineering*, 2007, pp. 34–43.
[52] E.M.D.B. Fávero, D. Casanova, and A.R. Pimentel, "SE3M: A model for software effort estimation using pre-trained embedding models," *Information and Software Technology*, Vol. 147, 2022, p. 106886.
[53] P. Liu, Y. Liu, X. Hou, Q. Li, and Z. Zhu, "A text clustering algorithm based on find of density peaks," in *7th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 2015, pp. 348–352.
[54] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
[55] W. Guohua and G. Yutian, "Using density peaks sentence clustering for update summary generation," in *Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2016, pp. 1–5.
[56] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
[57] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, Vol. 17, No. 11, 2015, pp. 1875–1886.
[58] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.
[59] B.A. Kitchenham, L.M. Pickard, S.G. MacDonell, and M.J. Shepperd, "What accuracy statistics really measure," *IEE Proceedings – Software*, Vol. 148, No. 3, 2001, pp. 81–85.
[60] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A simulation study of the model evaluation criterion MMRE," *IEEE Transactions on Software Engineering*, Vol. 29, No. 11, 2003, pp. 985–995.
[61] M. Korte and D. Port, "Confidence in software cost estimation results based on MMRE and PRED," in *Proceedings of the 4th International Workshop on Predictor Models in Software Engineering*, 2008, pp. 63–70.

[62] D. Port and M. Korte, "Comparative studies of the model evaluation criterions mmre and pred in software cost estimation research," in *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 2008, pp. 51–60.
[63] F. Sarro, A. Petrozziello, and M. Harman, "Multi-objective software effort estimation," in *38th International Conference on Software Engineering (ICSE)*. IEEE, 2016, pp. 619–630.
[64] T. Menzies, E. Kocaguneli, B. Turhan, L. Minku, and F. Peters, *Sharing data and models in software engineering*. Morgan Kaufmann, 2014.
[65] K. Muller, "Statistical power analysis for the behavioral sciences," 1989.
[66] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.
[67] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[68] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.
[69] A.M. Dai and Q.V. Le, "Semi-supervised sequence learning," *Advances in Neural Information Processing Systems*, Vol. 28, 2015.
[70] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian et al., "GPT understands, too," *arXiv preprint arXiv:2103.10385*, 2021.
[71] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov et al., "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, Vol. 32, 2019.

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230105, DOI: 10.37190/e-Inf230105

A Quality Assessment Instrument for Systematic Literature Reviews in Software Engineering

Muhammad Usman*, Nauman Bin Ali*, Claes Wohlin*
*Department of Software Engineering, Blekinge Institute of Technology, Sweden.
muhammad.usman@bth.se, nauman.ali@bth.se, claes.wohlin@bth.se

Abstract

Background: Systematic literature reviews (SLRs) have become a standard practice as part of software engineering (SE) research, although their quality varies. To build on the reviews, both for future research and industry practice, they need to be of high quality.
Aim: To assess the quality of SLRs in SE, we put forward an appraisal instrument for SLRs.
Method: A well-established appraisal instrument from research in healthcare was used as a starting point to develop the instrument. It is adapted to SE using guidelines, checklists, and experiences from SE. The first version was reviewed by four external experts on SLRs in SE and updated based on their feedback. To demonstrate its use, the updated version was also used by the authors to assess a sample of six selected systematic literature studies.
Results: The outcome of the research is an appraisal instrument for quality assessment of SLRs in SE. The instrument includes 15 items with different options to capture the quality. The instrument also supports consolidating the items into groups, which are then used to assess the overall quality of an SLR.
Conclusion: The presented instrument may be helpful support for an appraiser in assessing the quality of SLRs in SE.

Keywords: Systematic reviews, quality assessment, critical appraisal, AMSTAR 2, systematic literature review, tertiary study

1. Introduction

To establish evidence-based practices in software engineering (SE), Kitchenham [1] proposed the use of systematic literature reviews (SLRs) to identify, appraise and synthesise evidence reported in the scientific literature. Today the method is well-accepted in SE. This is illustrated by the growing number of published SLRs. The number of SLRs has increased rapidly since the introduction of the guidelines [2].

The reliability of conducting SLRs as a method needs to be maintained, ensuring trust in the results. Several researchers have proposed guidelines and checklists to design, conduct and report SLRs. However, relatively little work has been done on the critical appraisal of the SLRs. Several SE researchers have used an interpretation of the criteria used by the Centre for Reviews and Dissemination (CRD) at the University of York to include an SLR

Submitted: 16 Mar. 2022; Revised: 18 Jan. 2023; Accepted: 31 Jan. 2023; Available online: 09 Feb. 2023

in their Database of Abstracts of Reviews of Effects (DARE) [3]. However, these questions are insufficient to reveal significant limitations in SLRs [2, 4, 5]. These limitations act as a motivation for the research presented here.

In 2007, Kitchenham and Charters [6] introduced the DARE criteria [3] for quality assessment of SLRs in SE. The criteria come from healthcare and medicine. DARE included four criteria/questions. Since then, DARE has been updated to include five criteria/questions. According to Costal et al. [5], DARE is the most used quality assessment instrument of SLRs in SE, although many customize the criteria due to missing aspects in DARE or adapting it for their specific study.

According to Shea et al. [7]: *The Cochrane Collaboration Handbook provides a comprehensive guide for review authors, but it does not provide a concise critical appraisal instrument for completed reviews.* Thus, according to Shea et al. [7] the DARE criteria are primarily for authors of SLRs and not for assessing the quality of published reviews. According to CRD [8], the five criteria are used to decide whether to include an SLR in their database or not, which may be why it is not perceived to be sufficiently "concise and critical". It is sufficient that four out of five criteria are met for the SLR to be included in the CRD database. Thus, when combining the viewpoints of DARE, it implies that authors should ensure that they cover the DARE criteria for their SLR being accepted into the CRD database, i.e., the criteria are primarily inclusion/exclusion criteria. Thus, as noted, it is a quite high-level assessment, and it is not aimed at scrutinizing SLRs from a quality perspective. Despite the main objective of DARE, it has been used to assess the quality of published SLRs in SE. In a similar way as DARE is used in relation to the CRD database, DARE may be used as a screening instrument when deciding to include or exclude an SLR in a tertiary study. However, it is not suitable as a quality assessment instrument of SLRs. Costal et al. [5] conclude that there is a need for a comprehensive framework covering more aspects of quality than DARE.

To address quality assessment of published SLRs of randomized trials, AMSTAR was introduced in 2007 in medicine and healthcare. AMSTAR is, according to Shea et al. [7], one of the most used instruments for quality assessment. Through extensive use and validation of AMSTAR in healthcare and medicine [9], several necessary improvements were identified [10, 11]. AMSTAR 2 [7] was developed to address the limitations of AMSTAR, like handling non-randomized trials and being aligned with the revised Cochrane risk of bias instrument, which is described by Sterne et al. [12].

With the development of AMSTAR and its further improvement through AMSTAR 2, it is time to upgrade the quality assessment of published SLRs in SE instead of using the DARE criteria. Thus, we concur with Costal et al. [5] concerning the need for a more comprehensive framework covering more quality aspects, and on a more detailed level. We should continue to learn from other disciplines that are conducting SLRs. Therefore, building on the success of AMSTAR and its successor AMSTAR 2, we propose an adaptation of AMSTAR 2 to SE, which we call QAISER (**Q**uality **A**ssessment **I**nstrument for **S**oftware **E**ngineering systematic literature **R**eviews).

Our approach when developing QAISER has several salient features focusing on increasing the reliability of the research outcome. We based our work on a well-accepted and validated instrument as a foundation (i.e., AMSTAR 2) [7]. To ensure an appropriate adaptation to SE, we collected and relied on a comprehensive set of documents with guidelines and best practices for conducting SLRs in SE. We followed a systematic and well-documented process to develop QAISER with several internal validation steps involving multiple researchers. Furthermore, we invited some leading experts in evidence-based SE

research to conduct an external validation of QAISER. We also demonstrate the applicability of QAISER by using it to assess a sample of six selected systematic literature studies. In each step of the process, QAISER was updated based on the feedback received and internal discussions. The outcome of the process, i.e., QAISER, is the main contribution of the paper.

Our main objective is to support appraisers in assessing quality of completed SLRs, we believe that authors of SLRs may also use QAISER to help them with improving the quality of their SLR before submitting the research for assessment.

The remainder of the paper is organised as follows. Section 2 presents an overview of the main critical appraisal instruments used in both SE and evidence-based medicine. Section 3 describes in detail the method undertaken for developing QAISER. In Section 4, it is described how QAISER evolved into the latest version, which is the main outcome of the research presented. Section 5 describes how QAISER can be used to assess SLRs. In Section 6, we reflect on the reliability of QAISER. Section 7 describes the guidance document and shares our reflections about applying QAISER on six example SLRs. The threats to validity are presented in Section 8. Section 9 discusses the implication of the results. Section 10 concludes the paper, presents future research directions and our ambition to support broader adoption of QAISER. Finally, the QAISER instrument is provided in Appendix A (attached as supplemental material), and a guidance document supporting the instrument can be found in Appendix B (also attached as supplemental material).

2. Related work

A prerequisite for a quality appraisal is that we pose the right questions. In the first version of the guidelines for systematic literature reviews in SE, Kitchenham [1] identified two sets of questions from Greenhalgh [13] and Khan et al. [14] to review any existing SLRs on a topic of interest. In the 2007 update [6], Kitchenham and Charters added the CRD DARE set of four questions to the list [3]. Kitchenham and Charters [6] also applied the criteria to SLRs published between 2004 and 2007.

The proposal from Greenhalgh is very general; Khan et al. proposal is the most comprehensive, while the DARE criteria are brief and "simple" [6]. Among these three sets of questions proposed in the guidelines, only the DARE criteria have been widely used in the SE literature.

Kitchenham et al. [15] provided guidance to answer four of the five questions in the DARE criteria. Cruzes and Dybå [16] observed that one of the critical questions regarding synthesis had not been included in the SE guidelines for conducting SLRs and has not been used when evaluating the quality of SLRs in SE. It should be noted that the number of questions in DARE has varied over the years; it has included either four or five questions depending on the version of DARE.

Some others have developed their own interpretation of the DARE questions [17, 18]. One shared limitation of these is the lack of traceability between the proposals and the evidence/best practices used to motivate them.

Other researchers have also been concerned with assessing quality in SLRs in SE, Dybå and Dingsøyr [19] reviewed several proposals from evidence-based medicine to assess the quality of SLRs. They concluded that the MOOSE statement [20] is a very relevant reporting checklist for SLRs in SE. The MOOSE checklist has six main reporting items including

"background", "search strategy", "method", "results", "discussion" and "conclusions". Each item further lists actions and details that should be provided in an SLR.

In a previous study [21], we reviewed the proposals for quality assessment for SLRs both from SE and other fields. We concluded that in the SE literature, there is an awareness of reporting checklists like MOOSE, QUOROM, and PRISMA. However, SE researchers have not yet leveraged the progress in critical appraisal tools for systematic reviews.

One essential aspect related to quality assessment is the validity threats presented by authors of SLRs. Ampatzoglou et al. [22] reviewed 100 secondary studies in SE and identified the commonly reported threats to validity and the corresponding mitigation actions. They also proposed a checklist that authors can use to design an SLR with explicit consideration for common validity threats and develop an informed plan for mitigating them. The authors state that readers can also use the checklist to assess the validity of the results of an SLR. The checklist has 22 questions grouped into three categories: study selection validity, data validity, and research validity. Furthermore, for each of the 22 questions, there are 1 to 9 sub-questions.

The checklist by Ampatzoglou et al. [22] encapsulates the current state of research regarding mitigating validity threats. Also, the checklist is a useful design tool to support the design, execution and reporting of an SLR. However, we argue that it is not a tool that enables the evaluation of completed SLRs, e.g., should all items in the checklist be addressed? Even as a reporting checklist, Kitchenham et al. [23] point out the following major weaknesses in Ampatzoglou et al. approach and proposal: (1) they present what threats to validity are reported and not what should be reported, (2) they may have underestimated the extent of validity issues in secondary studies, and (3) they mix the threats to validity for mapping and reviews. Nevertheless, their work inspired the development of some QAISER items.

Table 1 presents an overview of various design and reporting checklists and assessment instruments for the quality assessment of systematic literature reviews.

Given the lack of an appraisal tool adapted for SE, we wanted to leverage experiences from other research fields. Through an analysis of the leading appraisal tools, including ROBIS, AMSTAR, and AMSTAR 2, we identified AMSTAR 2 (A MeaSurement Tool to Assess systematic Reviews) [7] as a candidate tool for adaptation to SE [21]. AMSTAR was developed based on a review of available rating instruments and consolidated them into 11 appraisal items. It has since been extensively used and validated. AMSTAR 2 is a revised version of the tool that takes into account the systematically collected community feedback. The major updates for AMSTAR 2 are: (1) the consideration of SLRs that may include non-randomized studies and (2) an increased focus on the risk of bias evaluation.

Table 1: An overview of checklists and assessment instruments for SLRs

Name	Approach	Domain	Awareness in SE
Ampatzoglou et al. [22]	Checklist	SE	Yes
MOOSE [20], QUOROM, PRISMA	Checklist	Medicine	Yes (see [19, 21])
PRISMA-ScR (for mapping studies), and ENTREQ and RAMESES (for qualitative reviews)	Checklist	Medicine	Yes (see [23]
SEGRESS [23]	Checklist	SE	Yes
DARE	Checklist	Medicine	Yes (see [4, 6])
DARE interpretation in SE (see [4, 6])	Assessment Instrument	SE	Yes
AMSTAR, AMSTAR 2, ROBIS	Assessment Instrument	Medicine	Yes (see [21])

AMSTAR 2 provides a more comprehensive coverage of important quality aspects of an SLR that are not included in the DARE criteria that are mostly used in SE [21]. AMSTAR 2 consists of 16 appraisal items and their corresponding response options and scale. Figure 1 annotates an example of an item, response, and scale from QAISER. QAISER kept the structure from AMSTAR 2. Nine QAISER items have three scale options (Yes, No, Partial Yes), while the rest have only two scale options (Yes, No). Like AMSTAR 2, QAISER includes "Partial Yes" in cases where it is relevant to recognise partial compliance with items. In the case of the six items with only two Yes/No options, partial compliance is not an option – i.e., for these items all response options are considered equally important. However, where possible, the alternate way to achieve a "Yes" rating is provided (Items 12, 13, and 15).

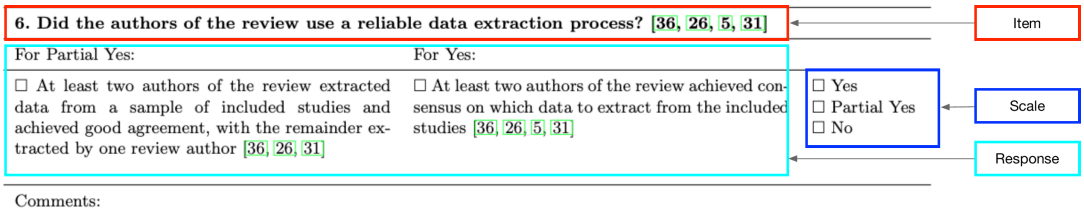

Figure 1: Items, responses and scale in QAISER

Based on an analysis of related work it was decided to use AMSTAR 2 as a basis for proposing a quality assessment instrument tailored for SE.

3. Method

This section describes the four-step process we used to develop QAISER (see Figure 2 for an overview). In the first step, we identified aspects from the evidence-based software engineering (EBSE) literature relevant for inclusion in QAISER. In the second step, we adapted AMSTAR 2 to SE by customizing its items and responses. In the third step, we combined the outputs of the previous two steps by integrating the EBSE aspects into QAISER. In the fourth step, we validated QAISER by inviting external experts to evaluate its completeness, understandability, and relevance of its items and responses for SE. Furthermore, we also used QAISER to assess a sample of six SLRs to demonstrate its applicability.

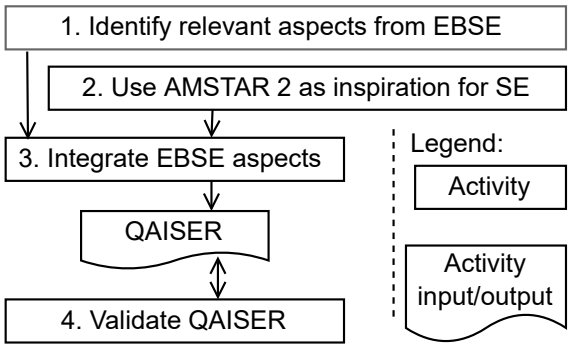

Figure 2: Overview of the QAISER development process

The first two authors jointly performed Steps 1–3 of the process, while the third author – the most experienced of the three authors – independently reviewed the work. Such division of roles among the authors was introduced to have an internal continuous sanity check on the outputs of all steps. Each step is further elaborated below and the details of each step are also illustrated in Figures 3–6 (the bidirectional arrows in these figures indicate that an activity results in the updates to its input).

Step 1: Identifying relevant aspects from the EBSE literature

In this step, we aimed to complement AMSTAR 2 with the relevant work from the EBSE literature. We followed a systematic approach to identify and analyze the relevant EBSE work (see Figure 3).

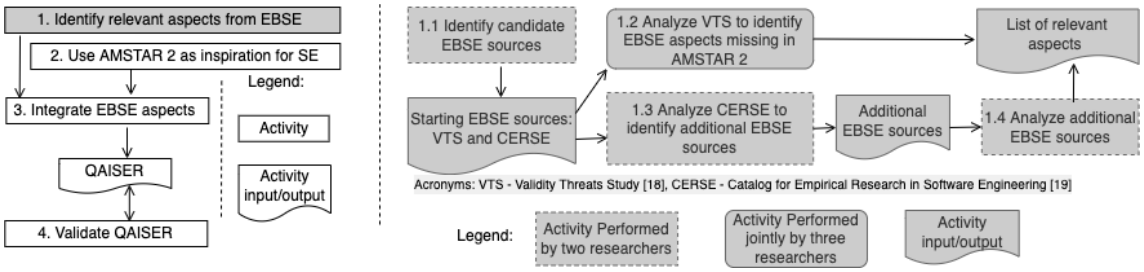

Figure 3: Step 1 – identifying relevant aspects from EBSE literature

We started with analyzing a closely related and recent tertiary study on validity threats in SLRs in SE by Ampatzoglou et al. [22]. They have aggregated validity threats and corresponding mitigating actions in the form of a checklist as described above in Section 2. We analyzed their checklist to identify aspects that are covered or missing in AMSTAR 2 [7].

Molléri et al. [24] recently proposed a Catalog for Empirical Research in Software Engineering (CERSE) based on a systematic mapping study of 341 methodological papers that were identified using a combination of manual and snowballing search strategies. CERSE includes available guidelines, assessment instruments, and knowledge organization systems for empirical research in SE. To identify additional relevant articles that are not covered by Ampatzoglou et al. [22] in their tertiary study, we selected 74 articles from CERSE that are related to SLRs and mapping studies (SMSs). We obtained the source file containing the basic information (title of the paper, publication venue, etc.) for these 74 articles from the first author of CERSE [24]. The first two authors independently reviewed these 74 articles to identify studies that propose or evaluate guidelines for conducting SLRs and SMSs in SE. Later, in a meeting, the first two authors developed a complete consensus on all 74 studies. The list of identified studies included, besides others, the latest version of the guidelines by Kitchenham et al. [25], the guidelines for mapping studies by Petersen et al. [26] and the guidelines for snowballing by Wohlin [27]. After including these three guidelines in our list of additional EBSE sources, we removed studies that were already covered in these guidelines [25–27].

Step 2: Using AMSTAR 2 as a source of inspiration for SE

The first two authors jointly analyzed AMSTAR 2 to identify items that are relevant for SE. As a validation, the third author independently reviewed the list of relevant and

non-relevant items identified by the first two authors. Next, the first two authors adapted the response options for SE, for example, by replacing the medicine-specific options with the appropriate SE options. The adapted response options were also reviewed independently by the third author. After discussions, we achieved complete consensus between all three authors on all changes in items and response options.

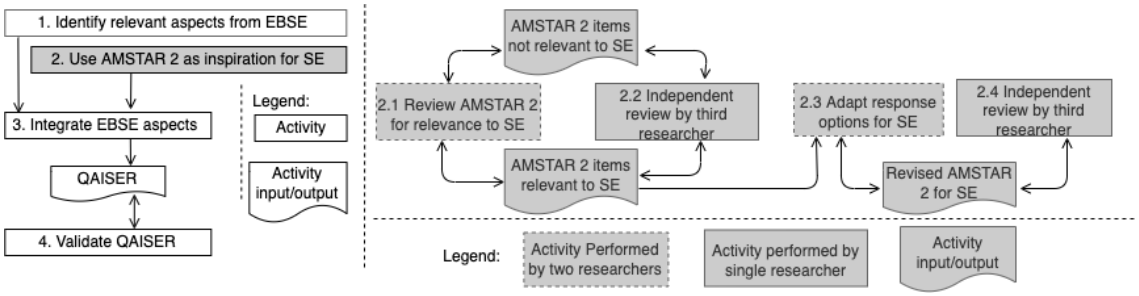

Figure 4: Step 2 – using AMSTAR 2 as a source of inspiration for SE

Step 3: Integrating EBSE aspects

Using the outputs of the previous steps and, in particular, the relevant EBSE literature identified in Step 1, the first two authors developed the first draft of QAISER. They also prepared a guidance document to support QAISER users in applying the instrument. The third author independently reviewed the instrument and the guidance document to validate its contents, i.e., to check that any relevant aspect is not missed. The independent review helped improve the formulations and remove some inconsistencies in the instrument and the guidance document. However, it did not result in any significant change in the instrument.

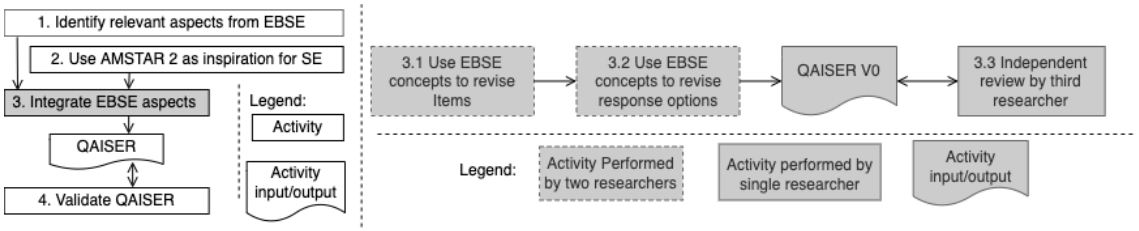

Figure 5: Step 3 – integrating EBSE aspects

Step 4: Validating QAISER

In this step, QAISER was reviewed by four experts in EBSE to validate the appropriateness of its items and reflect on its completeness (i.e., to identify if some aspects are missing) and understandability. The external experts are leading researchers in EBSE (see Table 2 for their profiles) and have been actively doing EBSE research since it was introduced in 2004 [28]. They have published several systematic secondary studies and made important methodological contributions to the EBSE discipline. In addition to QAISER and the guidance document, we prepared the following two documents to conduct the validation step (see Figure 6 for details about the validation step) systematically:

– A task description document: It described the steps that the external experts were asked to perform while reviewing QAISER. The task description document provided space where experts could enter their feedback on each QAISER item.
– A process description document: It briefly described the process we used to create QAISER.

Table 2: External experts' profiles

#	Published systematic secondary studies	Methodological contributions to EBSE	*h*-index*
Expert 1	Several	Yes – introduced SLRs to SE, several major contributions to the SLR guidelines	80
Expert 2	Several	Yes – several contributions, including mapping study guidelines	41
Expert 3	Several	Yes – contributions to a few specific steps/phases within the SLR process	38
Expert 4	Several	Yes – contributions to a specific step/phase within the SLR process	28

* At the time of the review.

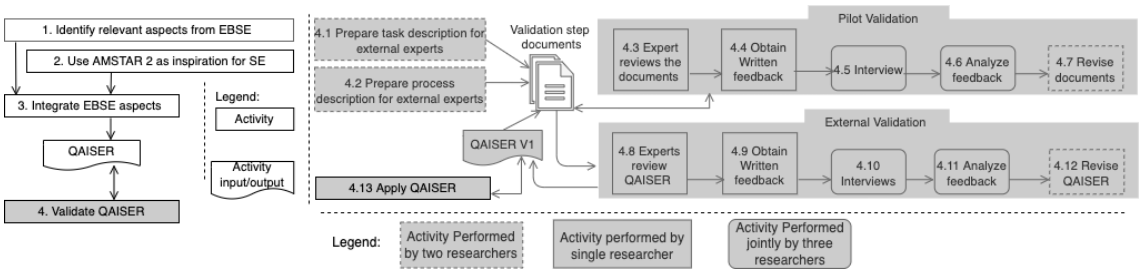

Figure 6: Step 4 – validating QAISER

Before the external validation, we performed a pilot validation with a senior colleague at our department who has experience of participating in multiple SLRs. The colleague reviewed all of the four documents mentioned above (i.e., task description, process description, QAISER, and the guidance document) and provided written feedback. We also conducted a follow-up interview (one hour, face-to-face) to discuss the feedback in detail and to ensure a shared understanding. We revised the task description and also the instrument based on the feedback collected during the pilot step. Most of the changes resulted in revised formulations. We shared the revised documents with our colleague and achieved consensus on the clarity of the task description and completeness and appropriateness of QAISER.

Next, we used the same approach with the external experts as we followed during the pilot. After obtaining the written feedback and performing the interviews (approximately one hour each and online) with all four external experts, we analyzed the comments to identify the changes that should be made in QAISER. Also, a revised version of QAISER and a summary of their feedback and our actions were sent to the external experts.

After the review of QAISER by external experts, we also applied it on a sample of six systematic literature studies (see Section 7 for more details, including how the six studies were selected). The application of QAISER resulted in some minor simplifications in the description of a few items. We also made our ratings of the six SLRs available online as additional support for using QAISER along with the guidance document and the QAISER

instrument as a spreadsheet. Section 7 shares our reflections about using QAISER as a critical appraisal tool, and the links where the application related material is available online.

4. Details of conducting the study

In this section, we present details of how we applied the process described in Section 3 and our justifications for the proposed changes to AMSTAR 2 while adapting it for SE.

In Section 4.1, we describe the development of the first version of QAISER (QAISER V0). This section is organized to explain the adaptations we made to AMSTAR 2 items, our justifications and the relevant EBSE sources used as the basis for the adaptations. For each AMSTAR 2 item, we start with the AMSTAR 2 item description, then provide our justifications for the proposed changes in the item, if any. QAISER V0 is not the final instrument. QAISER V0 was validated with the help of pilot and external validation steps. Section 4.2 presents the changes we made after the validation steps. In summary, Section 4.1 presents the output and the justifications of Steps 1–3 of the QAISER development process. Section 4.2 documents the changes to QAISER and their justifications made during Step 4 (validating QAISER) of the QAISER development process.

4.1. Development of QAISER V0

In Step 1 of the process described in Section 3, we identified and selected four sources (see [22, 25–27]), in addition to DARE [3], from the EBSE literature to identify the relevant aspects for QAISER. Later, based on the suggestions of the external experts, we also included two more sources for identifying the relevant aspects for QAISER. The two additional sources related to a framework for an automated-search strategy to improve the reliability of searches in SLRs [21], and a tertiary study describing lessons learned about reporting SLRs [4].

We now present the adaptation of AMSTAR 2 for SE based on the procedure detailed in Steps 2 and 3 of our method (see Section 3). Overall, at this stage in the process, we had two major changes in AMSTAR 2. The first change relates to the removal of existing items in AMSTAR 2. The removal includes excluding one item and replacing two items with a general item that is more appropriate for SE. The second change concerns the addition of an item.

In terms of removed items, three AMSTAR 2 items (Items 1, 11, and 12) were not included in QAISER as these were not deemed relevant to SLRs in SE. AMSTAR 2 Item 1 is about using PICO (Population, Intervention, Comparator group, Outcome) components in research questions and selection criteria. Items 11 and 12 are about meta-analysis, which is not commonly conducted in SE SLRs. We replaced these two items with a more general item about synthesis (see QAISER Item 11 in Appendix A). The new item checks if the included studies are synthesized or not. Synthesis of the included studies is one of the essential steps in an SLR [3, 4, 22, 25]. The details for these removed items are described later in the section.

The addition of one item is due to the following. Item 5 in AMSTAR 2 checks if the study selection is performed independently by at least two authors of the review. Item 6 checks the same aspect about the data extraction process. However, no item in AMSTAR 2 checks if the quality assessment is performed independently by at least two persons. We

introduced an additional item to cover this aspect, i.e., to see if the quality assessment is performed independently by at least two authors of the review (see QAISER Item 10 in Appendix A).

We now describe in detail why and what changes were made to each item in AMSTAR 2. For each item, we initially state its AMSTAR 2 formulation and then explain the changes we proposed in it.

Item 1. ***Did the research questions and inclusion criteria for the review include the components of PICO?***

The previous guidelines [6] suggested the use of PICO to structure the research questions. However, the revised guidelines [25] excluded the suggestion for using this structured approach for research questions. The guideline authors noted that for SE reviews the use of this structured approach has not been found useful due to the lack of consistent and stable terminology, which makes it hard to derive relevant search keywords [25]. Due to such issues, PICO has not been widely used in SE (for details see: [29, 30]). The issue of reporting inclusion criteria is discussed in the changes to AMSTAR Item 3.

Changes: This item is not relevant for SE and was excluded from QAISER.

Item 2. ***Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?***

We identified no need to make any change at the item level. However, the following issues in the response options were noted:

a) The response options for "Partial Yes" lack several aspects that are part of the protocol template included in the revised guidelines [25]. The missing aspects include description of the need for the review, data extraction process, synthesis process, threats to validity of the review, deviations from the protocol and the corresponding justifications for such deviations, and details of conducting the review.
b) Under "Partial Yes", the authors are only required to state that they had a written protocol. The authors should also make the protocol publicly accessible and describe where and how can it be accessed [25].
c) One of the response options uses the term "risk of bias assessment". In SE, the more commonly used term is quality assessment.

Changes: Based on the analysis, the response options were modified as follows as an adaptation of them for SE:

a) The missing response options under "Partial Yes" were added.
b) In the revised item, the authors are also required to make the protocol accessible and state how and where it can be accessed.
c) The risk of bias related response option was rephrased as quality assessment.

Item 3. ***Did the review authors explain their selection of the study designs for inclusion in the review?***

Most reviews in SE include different types of empirical studies. Thus, it is not relevant to ask for a justification for including all types of study designs. Furthermore, the study design is only one of the criteria for including or excluding studies from an SLR. Therefore, the item should address the larger aspect of the appropriateness of the inclusion and exclusion criteria. Reporting of the inclusion and exclusion criteria is also part of the DARE criteria [3] used by the Centre for Reviews and Dissemination at the University of York. Also, reporting of the inclusion and exclusion criteria and the relevant justifications is part of the guidelines [25] and other EBSE literature as well [22].

Changes: The item is reformulated as follows for SE: *Did authors of the review report*

their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?

To get a "Yes" score for the revised item, the review should have reported the inclusion and exclusion criteria and provided justifications for any restrictions used in the criteria.

Item 4. ***Did the review authors use a comprehensive literature search strategy?***

We identified the following issues in the response options:

a) The response options treat database search as the main search method while snowballing is partially addressed as an additional search method. In the revised guidelines for performing SLRs in SE [25], database and snowballing searches are included as alternate search strategies. Both strategies have been used in SE SLRs and have their own guidelines and best practices (for details see: [2, 25, 27]). In the current form of AMSTAR 2 Item 4 description, only the database search strategy could be assessed as comprehensive.
b) The response option related to the publication restrictions is more relevant to the inclusion and exclusion criteria.
c) Furthermore, two other response options are not used in SE SLR: The first one is about searching in the study registries, while the second one is about conducting the search within 24 months of completion of the review.

Changes: We introduced the following three changes:

a) Two groups of response options were created: first when a database search is used as the main search method and the second when a snowballing search is used as the main search method (See QAISER Item 4 in Appendix A) for details about the two groups of response options).
b) The response option related to the publication restrictions is moved to Item 3 (see Appendix A).
c) The two response options (searching in registries and search within last 24 months) were not included in QAISER. After the search is carried out, the review authors need to perform the remaining steps of the review process and write the review report. Finally, the review report is peer-reviewed before it can be published. The elapsed time in this entire process varies from case to case. To ensure the timeliness of the search, AMSTAR 2 includes this 24 months time limit, i.e., the search should have been conducted within the last 24 months. We removed the specific time limit of 24 months, which may not be appropriate in all cases. The appraisers are expected to judge the timeliness of the search (see Item 4 description in the guidance document in Appendix B) reported in the SLR under review.

Item 5. ***Did the review authors perform study selection in duplicate?***

We noted that:

a) The phrase "in duplicate" is not a commonly used term in SE and is therefore not self--explanatory. Furthermore, the item does not specify if the study selection is performed on the full text or on the titles and abstracts.
b) In the first response option, when all studies are reviewed independently by at least two authors of the review, the agreement level is not reported. Reporting of the agreement level would increase the transparency of the study selection process.
c) In the second response option, it is permitted that only a sample of the studies are independently reviewed by at least two authors of the review. The reliability of the study selection process is compromised if only a small sample of studies is reviewed by more than one author of the review. In particular, the excluded studies pose a threat to validity if a single person excludes them.

Changes: Three changes were introduced to address these observations:

a) The item was rephrased to clarify the focus on the independent study selection and that the initial study selection is based on titles and abstracts. The revised formulation is: *Did the authors of the review* ***independently*** *perform study selection* ***based on titles and abstracts****?*
b) At the end of the first response option, the following text is added to make it necessary to report the agreement level as well: *... and reported the agreement level.*
c) At the end of the second response option, the following text is added to make it compulsory to have the excluded studies reviewed by at least two authors: *however, all excluded studies must be reviewed by at least two authors of the review.*

Item 6. ***Did the review authors perform data extraction in duplicate?***

As in the previous item, the phrase "in duplicate" is not self-explanatory.

Changes: The item was rephrased in QAISER as follows: *Did at least two authors of the review independently perform data extraction?*

Item 7. ***Did the review authors provide a list of excluded studies and justify the exclusions?***

The item is about those studies that were excluded after reading the full text. The item does not indicate that it is about those studies that were read in full text, and not about those that were excluded based on the screening of the titles and abstracts.

Changes: The item was rephrased to indicate that it is about those studies that were read in full text. In the revised formulation, the following phrase is added at the end of the item text: *...for the papers read in full text?*

Item 8. ***Did the review authors describe the included studies in adequate detail?***

The response options about intervention and outcomes may not be relevant to all SLRs in SE. In SE, not all SLRs would be about interventions and outcomes. The included studies in an SLR may not have investigated any interventions. Furthermore, not all studies in SE include human subjects. In such studies, the population may consist of other relevant items of interest such as artifacts, events, or some other aspects. We have clarified it further in the guidance document (Appendix B) when describing Item 8.

Changes: In the response options about interventions and outcomes, the phrase *when applicable* is added to explain that the review needs to describe only the relevant information about included studies.

Item 9. ***Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?***

We noted the following:

a) Other fields use RoB assessments that focus on methodological rigor and the impact of weaknesses on the reliability of results. Kitchenham et al. [23] in a recent study, clarify the difference between RoB and quality assessment as follows: *"The important difference between RoB and quality assessment for individual studies is that RoB is about identifying potential methodological flaws that can bias the outcome of primary studies, whereas quality is about whether the research was performed as well as possible."*
   In SE SLRs, the concept of quality assessment is more prevalent than RoB. Therefore, in QAISER, we have used the term quality assessment.
   A variety of quality assessment instruments have been developed and used to assess the quality of the different types of empirical studies in SE [25]. The focus in SE so far has been on whether an SLR uses relevant quality assessment instruments. These instruments often cover both reporting quality and methodological rigor.

b) The current response options are not relevant to SE. Furthermore, the focus of the item is suggested to be changed to the quality assessment instrument. Therefore, the response options should also be revised accordingly to check the completeness and relevance of the questions in the quality assessment instrument.

Changes: We introduced the following changes:

a) We revised the item to emphasize whether or not the review authors have provided an explanation for their selection of the quality assessment instrument. The item is revised as follows: *Did the review authors explain their selection of quality assessment instrument?*

b) With regards to the response options under the revised item, for "Yes", the review authors should have selected an appropriate quality assessment instrument for different types of studies included in the review. Furthermore, the instrument needs to have questions about study goals, research questions, appropriateness of the study design, data collection, and analysis methods. The instrument should also have question(s) about the study findings and the supporting evidence, and the extent to which the findings answer the research questions. We refer to the instrument in Appendix A for the specific response options for this item in QAISER.

Item 10. ***Did the review authors report on the sources of funding for the studies included in the review?***

This item focuses only on the sources of funding for individual studies. Funding is one of the issues that could result in a conflict of interest. In some cases, the authors of the individual studies might have some other conflict of interest in favor of or against the topic or intervention they are investigating in their studies.

Changes: The item is revised to include any other conflict of interest besides funding sources. Conflict of interest is inserted in the item text as follows: *Did the review authors report on the sources of funding* ***and any other conflict of interest*** *for the studies included in the review?*

Item 11. ***If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?***

Meta-analysis studies are very rare in SE due to the lack of multiple empirical studies addressing the same research question. Therefore, this item is not relevant to the majority of the SE SLRs.

Changes: This item is removed from the adaptation of AMSTAR 2 for SE. We have instead included a more general item about synthesis (Item 11 in QAISER, see Appendix A).

Item 12. ***If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?***

As discussed with Item 11 above, meta-analysis is not common in SE SLRs. This item is removed from the adaptation of AMSTAR 2 for SE. However, it is important to note that considering the impact of the quality of individual studies while interpreting the results is still covered in the next item.

Item 13. ***Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review?***

We noted the following:

a) Instead of RoB, the SE community uses the notion of quality assessment more commonly.

b) The first response option deals with the inclusion of high-quality randomized controlled trials (RCTs). Since in SE, RCTs are not common, the focus should be on high-quality studies.
c) The second response option includes the requirement of discussing the impact of RoB on results. For SE, the focus has been on categorizing the analysis and interpretation of results based on study quality [25].

Changes: The following changes were introduced:

a) In line with Item 9 above, the RoB is replaced with quality of individual studies in the item description.
b) In the first response option, the phrase *high quality RCTs* is replaced with *high quality studies*.
c) The second response option is revised to focus on the categorization of the analysis and interpretation of results based on study quality.

Item 14. ***Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity[1] observed in the results of the review?***

We identified no need for adaptation to SE in this item.

Item 15. ***If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?***

a) The item is limited to quantitative synthesis. In SE, qualitative synthesis is used more frequently in SLRs. Discussing publication bias and its impact on the results is essential, regardless of the type of synthesis performed, quantitative or qualitative. Publication bias may be that some authors have contributed with several studies within the area of the SLR, and it may affect the conclusions from the SLR. The latter becomes particularly critical if it is one or more of the researchers conducting the SLR.
b) The response option includes a requirement to carry out graphical or statistical tests as well. The main aspect to cover in this item should be to check if the authors of the review have discussed publication bias and discussed its potential impact on review results.

Changes: We introduced the following changes:

a) The item is made more general by removing the word quantitative while also adapting its formulation for SE.
b) The response option is also revised accordingly, i.e., removing the reference to the graphical or statistical tests. The revised response option aims to check if the publication bias and its impact on the results are discussed or not.

Item 16. ***Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?***

We identified no need for adaptation to SE in this item.

We call the resulting instrument that systematically adapts AMSTAR 2 for SE and supplements it with SE guidelines and evidence QAISER V0. This version was used in Step 4 (see Section 3 for details of the process) for further validation.

4.2. Changes in QAISER during the validation step

This section presents the changes made in QAISER V0 based on the feedback collected during the pilot and external validation steps.

[1]Heterogeneity occurs when the results are not consistent across studies. For example, different studies provide conflicting evidence for or against a SE intervention. It is important to investigate the causes of such inconsistent results before drawing any conclusions in such cases.

Besides several editorial changes, the pilot validation resulted in the following two main changes in QAISER V0:

1) Addition of a new item on the need for undertaking the review (see QAISER Item 1 in Appendix A): In QAISER V0, establishing the need for undertaking a review was listed as one of the response options to score "Partial Yes" under Item 1. During the discussions in the pilot validation, we agreed with the senior colleague to give more importance to establishing the need for the review step. The number of SLRs performed in SE is increasing every year. At times, there are multiple SLRs on the same topic. Thus, there is a need to establish if it is relevant to undertake a new SLR on a topic [25, 26, 31]. The authors of the review should justify the need for undertaking the review. To score "Yes" on this new item in QAISER, the review should have 1) discussed the related existing reviews (if any) and established the need for another review by highlighting the gap, or 2) established the need to aggregate the evidence on a topic, if there exist no reviews on the topic.
2) Addition of a new response option under the synthesis related item in QAISER V0 (Item 11): Agreeing with the suggestion of the senior colleague, we added another response options under Item 11 in QAISER to check how effectively the authors of the review have linked the answers and interpretations with the data extracted from the primary studies. The new response option is described as: *Provided a clear trace linking the answers of review questions and interpretations to the data from the included primary studies.*

The revised QAISER, after the pilot validation step, was shared with the external experts for further validation. The external experts provided several improvement suggestions. We provide a summary of the main suggestions related to items and response options in the following:

- Introduce an item about recommendations: SLRs are supposed to provide evidence-based input to practitioners and researchers to aid them in making informed decisions. QAISER did not have any item that specifically covered this aspect. The external experts suggested including an item that checks if the review provides appropriate recommendations and conclusions based on the review results. Agreeing with the external reviewer's suggestion, we added a new item about recommendations and conclusions in QAISER (see QAISER Item 14 in Appendix A).
- Remove the item about sources of funding (see AMSTAR 2 Item 10 described in Section 4.1): The item deals with the reporting of the sources of funding for the included studies. The external experts suggested to remove it as they did not find it relevant in SE context. We removed this item from QAISER.
- Reformulate Items 5 and 6: The external experts suggested to reformulate QAISER Items 5 and 6 to describe them at an appropriate level. In the current formulations of Items 5 and 6, the requirement to include two authors was included both in the descriptions of the items as well as in response options. In line with the suggestion, we reformulated both items as: *Did the authors of the review include a reliable study selection (or data extraction) process?*.
- Introduce "Partial Yes" scale: Some items (Items 1, 5, 6, and 10) had a binary Yes/No scale. The external experts suggested introducing a third scale value of "Partial Yes" to make them more flexible. We introduced a "Partial Yes" option under these items and included the minimum acceptable requirements as response options (see QAISER Items 1, 5, 6, and 10 in Appendix A). AMSTAR 2 Items 5 (study selection) and 6 (data extraction) allow a "Yes" rating even if multiple researchers were involved in reviewing

only a sample of studies, with the rest being reviewed by only a single researcher. In the corresponding QAISER Items 5 and 6, such an approach (i.e., involvement of at least two researchers only on a sample of studies) results only in a "Partial Yes" rating. QAISER Items 5 and 6 have more stringent requirements of involving at least two researchers to review the eligible or included studies for a Yes rating (see QAISER Items 5 and 6 in Appendix A).

- Quality focus: Assessing SLRs is not only about the presence or absence of an aspect; it is largely a subjective judgment concerning decisions and measures taken by the authors. To incorporate this suggestion, we introduced adjectives such as adequately, reliable, and appropriate in several items to assess SLRs" subjective nature better.
- Modifications to the protocol-related item (see AMSTAR Item 2 described in Section 4.1): The external experts suggested simplifying the response options for the "Partial Yes" scale. We moved justification of any deviations from the protocol from "Partial Yes" to the "Yes" scale. Furthermore, threats to validity and details of conducting the review were removed from the "Partial Yes" scale. We also removed a response option about heterogeneity from the "Yes" scale. It was not deemed a necessary part of a protocol by the experts (see the revised description of QAISER Item 2 in Appendix A).
- Modifications to the heterogeneity-related item (see AMSTAR Item 14 described in Section 4.1): The external experts did not find this item to be essential for the systematic reviews in SE. The item is more relevant for meta-analysis studies, which are not common in SE. We replaced the heterogeneity concept with the characteristics of the primary studies. Some differences in the results of the primary studies may be due to the variations in the studies" characteristics, e.g., if the participants in different studies are students or practitioners. Therefore, in the case when there are differences in the results of the primary studies, the authors of the review should perform an analysis to see if the differences are due to the variations in the primary studies" characteristics.

4.3. Concluding remarks

QAISER aims to support appraisers of SLRs in SE by raising important questions about the reliability and the relevance of an SLR. Furthermore, by providing evidence-based and accepted best practices in SE research (i.e., established expectations in the SE field of a high quality SLR), it supports the judgement of the conformance and the likely impact of non-conformance on the reliability and relevance of an SLR.

The quality aspects of concern and related criteria in QAISER are based on available evidence and recommendations in the SE literature. Therefore, the availability of evidence and the specificity of guidelines is also reflected in the criteria used in QAISER. Thus, the responses in the instrument range from specific/concrete actions to broader/general suggestions/guidelines. QAISER supports appraisers in making a judgement about the overall reliability and relevance of an SLR.

5. QAISER as an appraisal instrument

QAISER has three levels of judgement: item level, group level, and SLR level. It should be noted that AMSTAR 2 does not include these three levels. The levels are introduced to support the appraiser in moving towards an overall assessment of an SLR. However, the levels do not imply that QAISER aggregate the overall assessment to a final numeric score.

The use of a single aggregate numeric score to compute and reflect on the quality of an SLR is not a recommended practice anymore [7] – instead a subjective assessment on an ordinal scale is preferred (e.g., high, medium, low, ad critically low ratings in AMSTAR 2 for the overall confidence in review results).

In this section, the three levels are presented in Section 5.1 (item level), Section 5.2 (group level) and Section 5.3 (SLR level) respectively.

Some items are more closely related to each other, e.g., Items 3, 4, 5, and 7 relate to the identification and selection of potentially relevant studies. Therefore, to allow appraisers to reflect on the strengths and weaknesses of the SLR in a group of related items in one place, we introduced the concept of group level assessment. After performing the item level assessment, appraisers perform the group level assessment, allowing them to assess the SLR on a group of related items. After this group level assessment, appraisers consolidate their assessment at the overall SLR level to judge if the SLR, as a whole, is reliable and relevant. At the SLR level, the assessments in the related groups support judging the relevance (two groups related the relevance: Groups 1 and 6) and reliability of the SLR (five groups related to the reliability: Groups 2, 3, 4, 5, and 7). Table 3 presents the items and the groups of QAISER, while the complete instrument and the guidance document are presented in Appendix A and B respectively (see supplemental material).

5.1. QAISER: item level assessment

The first level comprises 15 items formulated as questions. These questions are ordered to reflect the sequence of phases in the design, conduct, and reporting of a typical SLR. The criteria to meet the questions on the item level are stated in the form of acceptable responses for each of the questions. All items are evaluated on a scale with two values (Yes/No) or three values (Yes/Partial Yes/No), i.e., an assessment of the extent to which an SLR under review fulfils the stated criteria.

Each item in QAISER is formulated with the objective that it is self-contained and self-explanatory. However, there is an accompanying guidance document (Appendix B in the supplemental material) with a more detailed description of the items and their responses. We recommend that before applying QAISER, the guidance document should be read, at least, before using QAISER for the first time.

5.2. QAISER: group level assessment

The external experts also provided a suggestion about clarifying the flow and sequence of the items in QAISER. To make the flow of the items more explicit and understandable, and to aggregate individual items into a logical cluster, we organized the 15 QAISER items into seven groups corresponding to the process and outcome of an SLR (see the first column in Table 3): (1) motivation to conduct a review, (2) plan and its validation, (3) identification and selection, (4) data collection and quality appraisal, (5) synthesis, (6) recommendations and conclusions, and (7) conflict of interest.

At the group level, the assessment results on the item level are used as indicators for major and minor weaknesses based on their impact on the reliability and relevance of an SLR, see Table 4. Having completed the assessment of individual QAISER items, an appraiser should reflect on the impact of the weaknesses on the reliability and relevance of the SLR at the group level. Groups 1, 2, 6, and 7 consist of single items only, and are therefore relatively simple to reflect upon. A “No” rating on the corresponding items of

Table 3: QAISER items and groups

Group	Item description and the relevant sources/references
1. Motivation	Item 1: Did the authors of the review adequately justify the need for undertaking the review? [22, 25, 26]
2. Plan	Item 2: Did the authors of the review establish a protocol prior to the conduct of the review? [7, 22, 25]
3. Identification and selection	Item 3: Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions? [3, 22, 25] Item 4: Did the authors of the review use a comprehensive literature search strategy? [3, 7, 22, 25] Item 5: Did the authors of the review use a reliable study selection process? [7, 25, 26] Item 7: Did the authors of the review discuss and justify the exclusion of the potentially relevant studies that were read in full text? [7, 25]
4. Data collection and appraisal	Item 6: Did the authors of the review use a reliable data extraction process? [7, 22, 25, 26] Item 8: Did the authors of the review provide sufficient primary studies" characteristics to interpret the results? [3, 7, 22, 25] Item 9: Did the authors of the review use an appropriate instrument for assessing the quality of primary studies that were included in the review? [7, 25] Item 10: Did the authors of the review use a reliable quality assessment process? [25]
5. Synthesis	Item 11: Were the primary studies appropriately synthesized? [3, 4, 22, 25] Item 12: Did the authors of the review investigate the impact of the quality of individual studies on the results of the review? [7, 22, 25] Item 13: Did the authors of the review investigate the impact of primary studies" characteristics on the results of the review? [7, 25]
6. Recommendations and conclusions	Item 14: Did the authors of the review provide appropriate recommendations and conclusions from the review? [4]
7. Conflict of interest	Item 15: Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review? [7, 25]

these four groups indicates a major weakness at the group level. Groups 3, 4, and 5 consist of multiple items and are more complex to reflect upon. The appraisers should make an overall assessment after considering the ratings of all items in the groups. As a rule of thumb, we recommend that all items receiving a "No" should be considered as hinting at a major weakness in the group being assessed.

5.3. QAISER: SLR level assessment

By progressively building on the first two levels, an appraiser judges the overall reliability and relevance of an SLR at the SLR level. Thus, considering the impact of weaknesses in related groups, i.e., **relevance** (mainly two groups: motivation, and recommendations and conclusions) and **reliability** (mainly the following five groups: plan, identification and selection, data collection and appraisal, synthesis, and conflict of interest).

AMSTAR 2 suggests that the appraisers should pre-specify which items are more critical for an SLR under review. They have also suggested an advisory list of critical items. For QAISER, we recommend a similar approach. The appraisers should pre-specify which groups of items are more/less critical for the specific review being assessed.

Table 4: QAISER: group level assessment

Group	Item ranking (Yes/Partial Yes/No)	Impact	Comments
1. Motivation	Item 1 (need):		
2. Plan	Item 2 (protocol):		
3. Identification and selection	Item 3 (selection criteria): Item 4 (search): Item 5 (selection process): Item 7 (excluded studies):		
4. Data collection and appraisal	Item 6 (data extraction): Item 8 (study characteristics): Item 9 (quality criteria): Item 10 (quality assessment process):		
5. Synthesis	Item 11 (synthesis): Item 12 (considered study quality) Item 13 (considered study characteristics)		
6. Recommendations and conclusions	Item 14 (recommendation):		
7. Conflict of interest	Item 15 (their own):		

In the following text, we provide two examples of using the group-level assessment for assessing the reliability and relevance of an SLR.

Interpretation criteria, example 1:

1) Reliability of an SLR: The reliability of an SLR is assessed by rating the overall confidence in the results of an SLR as: high, moderate, low or critically low. Apart from group 1, all other groups are relevant while considering the confidence in the results of an SLR. As a rule of thumb, we recommend that the confidence in the SLRs with major weaknesses in groups 3, 4, and 5 should be considered "critically low".
   Table 5 provides guidance for interpreting weaknesses observed at the group level to select a confidence rating at the SLR level.

Table 5: Judging the reliability of the results of an SLR

Reliability	Suggestions for interpretation based on group-level assessment in Table 4
"critically low"	– major weaknesses in groups 3, 4, and 5
"low"	– major weaknesses in at most two of the groups 3, 4, 5 along with major weaknesses in groups 2 and 6
"moderate"	– no major weakness in groups 3, 4, 5, and 7, but major weaknesses in groups 2 or 6
"high"	– only minor weaknesses in at most two of the groups 3, 4, 5 and only a few minor weaknesses in groups 2, 6, and 7

2) Relevance of an SLR: The relevance of an SLR is also rated as high, moderate, low or critically low. Groups 1 and 6 are considered when making a judgement about the relevance of an SLR. As a rule of thumb, we recommend that the relevance of an SLR be judged to be "critically low" if there are major weaknesses in both groups 1 and 6. Table 6 provides guidance in selecting a relevance rating based on the weaknesses in groups 1 and 6.

Interpretation criteria, example 2: A more stringent criteria for assessing reliability (instead of Table 5) could be stated as follows:

– Critically low – major weakness in at least one of the groups 3, 4, 5, or 7.

Table 6: Judging the relevance of an SLR

Relevance	Suggestions for interpretation based on group-level assessment in Table 4
"critically low"	– major weaknesses in group 1 and 6
"low"	– major weakness in either group 1 or 6
"moderate"	– minor weaknesses in both groups 1 and 6
"high"	– only a minor weakness in group 6

– Low – several minor weaknesses in groups 3 and 4, and 5.
– Moderate – no major weakness in groups 3, 4, 5, and 7, but a major weakness in groups 2 or 6.
– High – only minor weaknesses in at most two of the groups 3, 4, 5 and only a few minor weaknesses in groups 2, 6, and 7.

The above criteria acknowledge that a major limitation in any one of the groups 3, 4, 5, or 7, i.e., search and selection (group 3), data collection and appraisal (group 4), synthesis (group 5), or conflict of interest (group 7) cannot be compensated for by excellence in other groups. For example, a thorough synthesis does not compensate for limitations in the search.

6. Reliability of QAISER

In this section, we highlight three aspects that contribute to the reliability of QAISER as a potentially effective instrument for assessing the quality of SLRs in SE.

1) The relevance of AMSTAR and AMSTAR 2 validations: The original AMSTAR [32] consisted of 11 appraisal items. Based on community feedback, AMSTAR 2 was proposed consisting of 16 items with an increased focus on the risk of bias evaluation and the possibility to assess SLRs that may have non-randomized studies. Both AMSTAR and AMSTAR 2 have been used and validated extensively (for details see:[9, 33, 34]). These validation efforts provide credibility to QAISER as well, as most of its items (12 out of 15) are adapted from AMSTAR 2.
2) Comparison with DARE: DARE [3] is the most frequently used criteria to assess the quality of SLRs. Several essential aspects related to the quality of SLRs are not covered in DARE, e.g., justifying the need to conduct a review, establishing a protocol prior to performing the review, study selection process, data extraction process, and quality assessment process. Furthermore, three of the DARE criteria (including the important criterion about synthesis) are limited to checking the presence/absence of different aspects, rather than their appropriateness, e.g., *if the inclusion and exclusion criteria are reported or not?* QAISER not only covers aspects that are missing in DARE, but it also focuses on quality aspects of different criteria – for example, checking the appropriateness of inclusion and exclusion criteria rather than only focusing on the mere reporting of such criteria in the review report.
3) External validation. We followed a systematic process (see Section 3 for details) to adapt AMSTAR 2 for SE and to introduce additional aspects based on the recommendations in the systematically identified EBSE literature. Four leading experts then reviewed the proposed instrument to check the appropriateness, completeness, and understanding of its items (refer to Section 3 for details about the validation step). The experts recommended some changes, which we incorporated in the revised version of QAISER.

The experts did not suggest any further changes in the revised version. Although QAISER is developed using a systematic process, the proposed changes while adapting AMSTAR 2 to SE still need to be empirically validated by independent researchers.

7. Application of QAISER

In this section, we describe the support available for applying QAISER and share our reflections on applying QAISER to assess a sample of six systematic literature studies. The selected studies include five systematic literature reviews and one mapping study [35] published in SE. None of the five SLRs used meta-analysis, and hence the adaptations made in QAISER for assessing more qualitative SLRs were tested. For example, [36] aimed at meta-analysis, but had to settle for vote counting, and [37] used three synthesis approaches: narrative synthesis, vote counting and reciprocal translation. Thus, the analysis includes two approaches being qualitative of nature. The other three systematic literature studies, [38, 39] and [40] also used more qualitative approaches to synthesize their findings, including, an adaptation of comparative analysis, and in the other two SLRs, categorisation was conducted by the authors. Thus, the five SLRs primarily use qualitative approaches to synthesize the findings.

7.1. Support for applying QAISER

In line with AMSTAR 2, we also developed a guidance document (see Appendix B) for supporting appraisers in applying QAISER. The guidance document describes the following aspects for each QAISER item:

1) What is the item about? We provide a brief description of the item.
2) How to assess the item? We explain what an appraiser needs to check for assigning "Partial Yes", "Yes", and "No" ratings.
3) Where to find the relevant information to assess the item? We provide hints concerning which sections of the review papers are most likely to have the information needed to assess the item.

To further support the application of QAISER, we developed a spreadsheet to operationalize the QAISER instrument.

7.2. Reflections on Applying QAISER

We applied QAISER on a sample of six systematic literature studies (SLS) [35–40] – five SLRs and one SMS. The six systematic studies were selected from four tertiary studies. The four tertiary studies were selected based on being published recently (2018-2020) and using DARE with five criteria. The purpose of applying QAISER on a sample of SLSs was to validate its usefulness – and also to make available the ratings of the selected SLSs as additional support for applying QAISER. The six SLSs were selected as follows:

– The second author selected three SLRs from the tertiary study by Kitchenham et al. [4] that were previously assessed using DARE with the highest ranking. The purpose of selecting the high-ranking SLRs on DARE criteria was to illustrate the usefulness of QAISER in supporting appraisers in performing a more fine-grained and thorough critical appraisal compared to DARE.

– The third author selected three SLSs, one from each of the other three tertiary studies meeting the criteria concerning publication year and using DARE with five criteria. One of the tertiary studies is also generic in a similar way as the tertiary study above [41]. However, it covers mapping studies. It was decided to include this tertiary study to evaluate DARE vs QAISER on a mapping study too. The other two tertiary studies cover different topics within SE [42, 43]. For each of these three tertiary studies, one systematic study was randomly selected. The DARE scores for the three selected SLSs were 2, 2.5 and 3, respectively. Thus, the objective was to also apply QAISER on SLSs with lower DARE scores than the highest score (5), i.e., to contrast with the three SLRs selected by the second author.

We assessed the six selected SLSs with QAISER in two pairs of raters: 1) the first and the second author for the three SLSs selected by the third author, and 2) the first and the third authors for the three SLRs selected by the second author. Each QAISER item was used six times – on six different SLSs. The assessment resulted in the application of 15 QAISER items on six selected SLSs by the two pairs of raters, i.e., in total 90 assessment decisions for six SLSs. For a majority of the items, there were either no or minor differences (11 minor differences in total of the type Yes/Partial-Yes or Partial-Yes/No). There were also 12 differences of the type Yes/No in total – for no item we had more than two major differences of Yes/No type. Looking at both types of differences – minor and major – we noticed that for Item 1 (motivation), Item 2 (protocol) and 14 (Recommendations), we had relatively more differences as compared to the other items. Nevertheless, all authors found the option to add comments extremely helpful in reflecting on our ratings and corresponding justifications. Besides sharing raters' perspective, the comment fields also provide an opportunity to reflect on those aspects that are not captured in the QAISER instrument.

We also noted that the possibility to build on the item level assessment at the group and overall SLR level helped us in arriving at an overall rating for the SLR in a subjective, but informed and structured way. Both pairs of raters arrived at similar ratings at the SLR level, i.e., there were no major disagreements. For the relevance judgement, three SLSs received exactly the same rating, while the remaining three had minor differences as follows: high/moderate ratings (two SLSs) and low/moderate ratings (one SLS). Likewise for the confidence judgment as well, three SLSs received exactly the same ratings, while the remaining three have minor differences as follows: low/critically low ratings (two SLSs), low/moderate ratings (one SLS).

We observed that the time spent on QAISER application is relatively more when it is used for the first time. But for the second and third application, all authors noted that we were able to complete the assessment relatively quickly. In case of the tertiary studies wherein the quality of several SLRs needs to be assessed, the authors would have to spend some time in the start – on studying the QAISER instrument and the guidance document before the first time use – but can then continue to use it on the remaining SLRs more efficiently. Using QAISER instead of DARE would relatively be more time-consuming. However, QAISER provides support in performing a more fine-grained and thorough critical appraisal compared to DARE. Even for those aspects that are already covered in DARE (e.g., inclusion and exclusion criteria and search strategy), we were able to perform a more fine-grained appraisal. The support of QAISER worked equally well in assessing the systematic mapping study with one expected difference related to the time spent on the assessment. The assessment of the mapping study [40] with QAISER was relatively

less time-consuming as we did not have to apply the QAISER items related to the quality assessment and synthesis (i.e., Items 9–12).

The guidance document, QAISER instrument, and the spreadsheet corresponding to the six example applications are all available online[2]. Researchers could look at the six examples as additional support complementing the guidance document.

Based on our experiences of applying QAISER on the selected systematic studies, we decided to make the following minor simplifications in the QAISER instrument:

- Item 2 (Protocol): In the previous version, there was a requirement that the protocol should be publicly available. For the six SLSs we assessed, we noted that most authors included the necessary parts of the protocol in the review report. We made a small change in the instrument that allows authors to either include the protocol in the review report or make it available online.
- Item 15 (Conflict of interest): In the previous version, this item explicitly covered reporting of the funding sources and any other potential sources of conflict of interest. The guidance document covered the possible impact of the review authors" own prior research on the review results while explaining "any other potential sources of conflict of interest". We noted that financial conflicts of interest were irrelevant for most studies during the assessment of the six selected SLSs. However, the potential impact of review authors" own prior research on the review results became apparent. Therefore, we made a minor change in this item by lifting this consideration from the guidance document and making it part of it. The item scale has been slightly adjusted to reflect the renewed focus.

Some QAISER items and response options were also reformulated during the peer review process based on the recommendations of the reviewers. In addition, the review process also resulted in a few minor changes in the guidance document. The changes in the descriptions of the QAISER items during the review process include the following:

- To link the inclusion and exclusion criteria of the review with its research questions, Item 3 is revised as: *Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?*
- To further clarify the expectation from the authors of the review to analyze the impact of the quality of the included studies on the review results, Item 12 is slightly modified as: *Did the authors of the review investigate the impact of the quality of individual studies on the results of the review?*
- To further clarify the expectation from the authors of the review to analyze the impact of the characteristics of the included primary studies on review results, Item 13 is also reformulated: *Did the authors of the review investigate the impact of primary studies characteristics on the results of the review?*
- Item 15 is also simplified as: *Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?*

The final version of the QAISER is presented in Appendix A.

8. Threats to validity

In this research, we aimed to propose an instrument for appraising systematic literature reviews in SE. In the design and conduct of this research, the following two objectives guided us:

[2]https://drive.google.com/drive/folders/1p7OUEfqQTF4dY3e_OX_OHiyi_tC4E_cU?usp=sharing

1) To develop an instrument that is comprehensive, practical and appropriate for SE. Thus, the instrument shall cover all essential elements concerning the quality of an SLR, assist the appraiser when judging the quality of an SLR, and take into account the SE body of knowledge.
2) To reduce the researcher's bias in the development of the instrument.

The two main threats to validity identified concerned researcher bias and applicability of QAISER. The researchers come from the same affiliation, which creates a risk of having a coherent view on research. When creating an instrument for use by the research community, there is a risk that the instrument is hard to understand, and hence limiting its applicability.

To achieve the two objectives above and mitigate the threats to validity, we undertook the following actions:

- **Benefits and limitations of using AMSTAR 2 as a foundation for QAISER:** We used AMSTAR 2 as the starting point for our work as it is a well used and validated tool [21]. Albeit, AMSTAR 2 is primarily developed for quantitative studies as shown through Item 11 and 12, which focus on meta-analysis. However, as qualitative studies and mapping studies are more common in SE than quantitative studies using meta-analysis, we have adapted such items from AMSTAR 2 in QAISER to match the software engineering context. The adaptation was assessed through the external experts and applying QAISER to five SLRs and one mapping study from SE. None of the six studies used meta-analysis. As a consequence, items inherited from AMSTAR 2 have been validated to a larger extent than the newly introduced items in QAISER (Items 1, 10, and 14) and the adaptations to SE.
- **A systematic and rigorous process:** As described in Section 3, we followed a systematic approach for the development of QAISER. All data collection, analysis and interpretation involved at least two researchers. A third researcher independently reviewed the outcomes from several individual phases in the study. We maintain traceability for the adaptations in the existing tool by documenting the reasons, sources consulted and the changes.
- **Validation with leading experts in the field:** QAISER was reviewed by only four external experts, which is a limitation of our study. The number of experts is limited to four, but they include some of the main contributors of the methodological guidelines for designing, conducting and reporting secondary studies in SE. They have also authored several SLRs and have conducted several studies reporting critical evaluations of existing SLRs in SE.
- **Capturing agreed design, process and reporting guidelines in QAISER:** QAISER aims to support appraisers in assessing the quality of completed SLRs. As a critical assessment instrument, it ought to reflect the current best practices and established guidelines regarding the systematic review design, process, and reporting in software engineering. For this purpose, we used a systematic approach to identify and select a representative and current list of sources to capture the current best practices and evidence for inclusion in QAISER (see Step 1 in Section 3 for details of our approach and the sources considered when developing QAISER). Thus, we minimized the risk of overlooking relevant literature or proposing an instrument in conflict with the current best practices by undertaking the following actions: (1) a thorough search utilizing two recent and comprehensive reviews of methodological guidelines [22, 24], (2) the inclusion of key guidelines [25, 26], and (3) consulting experts who are the main contributors to the methodological guidelines in software engineering.

With these actions, we have tried to mitigate the major threats to the validity of QAISER. However, the instrument needs to be used by other researchers beyond the authors of QAISER. Similar use and validation of AMSTAR identified several necessary improvements and resulted in AMSTAR 2. Such an evaluation of the instrument is essential to form a basis for further improvement of QAISER related to aspects like usability and reliability. QAISER introduced the concept of item, group, and SLR level assessments. This multi-level assessment was introduced to support appraisers in arriving at an overall assessment at the SLR level. Apart from the six SLSs assessed by the authors, this multi-level assessment mechanism has not been validated by external researchers. We plan to conduct external validation as part of our future research. We have made QAISER publicly available on online research forums[3]. In this paper, the idea is to publish QAISER with the current validation – wherein QAISER has been reviewed by a few experts (during pilot testing and external review) and applied by the authors to a sample of SLRs – and then continue to update it based on the community feedback and evidence from the future external validation studies.

By making QAISER and guidance for its usage publicly available, we hope that we and others in the field will address this need in the future.

9. Discussion

Given the lack of an appraisal instrument for assessing the quality of SLRs in SE, we developed QAISER. As presented in the introduction (see Section 1), researchers in SE have used the criteria in DARE [3] for assessing the quality of SLRs, although it comes with limitations [2]. Furthermore, to simply use an appraisal instrument, such as AMSTAR 2, from another discipline also comes with issues as illustrated in the development of QAISER. There was a need to adapt AMSTAR 2 to SE, and hence AMSTAR 2 is not an option by itself. The differences between disciplines need to be captured in the appraisal instrument.

QAISER takes its starting point from a well-established appraisal instrument from another field, i.e., AMSTAR 2 from the field of evidence-based healthcare. Furthermore, QAISER incorporates best practices from the conduct of SLRs in SE. Thus, QAISER is well-grounded in the literature and contributes to taking the quality assessment of SLRs in SE one step forward.

The objective of QAISER is to support appraisers of SLRs in SE. The expertise of the individual appraisers is crucial, and it cannot be replaced with an appraisal instrument such as QAISER.

The DARE assessment criteria are used to aggregate the ratings on the five questions as a final numeric quality score (maximum value is five – corresponding to five DARE questions). QAISER, like AMSTAR 2, does not assign a final numeric score for quantifying the quality of an SLR. On the contrary, QAISER is intended to support appraisers by covering the most salient quality aspects of an SLR. Thus, QAISER will help identify major and minor weaknesses in the design, execution, or reporting of an SLR that compromise the SLR results" relevance and our confidence in them.

Although the main objective is to support appraisers in assessing the quality of completed SLRs, we believe that authors of SLRs may also use QAISER to help them with improving

[3]QAISER is publicly available on ResearchGate https://www.researchgate.net/publication/354765980_A_Quality_Assessment_Instrument_for_Systematic_Literature_Reviews_in_Software_Engineering and on arXiv https://arxiv.org/abs/2109.10134

the quality of their SLR before submitting the research for assessment. In the best of worlds, each submitted SLR is of high quality at the submission stage. It should be noted that the quality of an SLR is highly influenced by the quality of the primary studies included. The need to assess the quality of primary studies is highlighted by, for example, Dybå and Dingsøyr [19], and Yang et al. [44]. With the same objective, Wohlin highlights the need to write for synthesis when publishing primary studies [45].

We recommend all users of QAISER look at not only the appraisal instrument itself but also the accompanying guidance document. The latter is particularly important when using the instrument for the first couple of times. We have also made available online a spreadsheet operationalizing QAISER and six example assessments to further support appraisers in using QAISER.

The items and their response options in QAISER are intended to help highlight areas with weaknesses (or room for improvement). Given that assessment is prone to bias, we have deliberately chosen to have two or three levels for assessing each item. More levels may increase the risk for appraiser bias, although it may also benefit since the scale becomes more fine-grained. However, since QAISER is geared towards supporting appraisers of SLRs, we leave it to each appraiser to tune the feedback in writing using the comments option provided with each item, rather than having a more fine-grained scale.

When using QAISER for a mapping study, some items or questions may be less applicable than for an SLR, for example, the item concerning synthesis. We did consider adding an option of "not applicable" for mapping studies. We have chosen not to make the appraisal instrument more complex by adding the "not applicable" option. Thus, we leave it to each appraiser to decide if something is not applicable for a mapping study. Our preference is to leave freedom to the appraiser, given that SLRs and mapping studies may come in different shapes and colors. Assessing SLRs and mapping studies is a subjective endeavour, and the objective of any appraisal instrument should be to support the expert appraiser.

10. Conclusion and future work

QAISER, as an appraisal instrument for SLRs in SE, is built on a well-established appraisal instrument from another discipline (AMSTAR 2), and a set of guidelines, checklists, and experiences from SE. Furthermore, four external experts on SLRs in SE have reviewed an earlier version of QAISER, and QAISER has been revised based on their feedback. QAISER has also been used to assess the quality of six selected SLRs to demonstrate its applicability. Thus, QAISER is well-founded, and hence it is ready for further validation through usage.

QAISER includes 15 items and several response options for each item to assess for appraisers to arrive at an assessment for each item. QAISER provides support to consolidate the items on a group level, which is not done in AMSTAR 2. In QAISER, the items are consolidated into seven groups to support the appraiser to get a good overview of the strengths and potential weaknesses of an SLR. Moreover, QAISER has support for consolidating from the group level to the SLR level. The assessment of each group is systematically used to form an opinion about the overall quality of an SLR both in terms of reliability and relevance. AMSTAR 2 only provides an overall assessment of the confidence in the results. Given the importance of both reliability and relevance of the results for SE, we have provided support for both aspects.

In the future, we plan to evaluate the reliability and usability of QAISER by asking independent researchers to use it to assess the quality of selected SLRs. Based on such feedback, we plan to enhance QAISER further to support the SE community in assessing SLRs.

Acknowledgements

We would like to express our sincere thanks to the external experts: Prof. Daniela S. Cruzes, Prof. Barbara Kitchenham, Prof. Stephen G. MacDonell and Prof. Kai Petersen for their constructive feedback on QAISER.

We would also like to thank Prof. Jürgen Börstler for his kind participation in the pilot of the study. His detailed feedback helped us to improve the planning and execution of the evaluations with external experts. We also extend our gratitude to Dr. Jefferson S. Molléri for providing the listing of articles from the work with CERSE. Lastly, we are also thankful to the anonymous reviewers for their value feedback, which has helped us to further improve the Manuscript.

This work has been supported by ELLIIT, a Strategic Area within IT and Mobile Communications, funded by the Swedish Government. The work has also been supported by research grants for the VITS project (reference number 20180127) and the OSIR project (reference number 20190081) from the Knowledge Foundation in Sweden.

References

[1] B. Kitchenham, "Procedures for performing systematic reviews," Keele University, Keele, UK, Tech. Rep., 2004.

[2] N.B. Ali and M. Usman, "Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy," *Information and Software Technology*, Vol. 99, 2018, pp. 133–147.

[3] Centre for Reviews and Dissemination, University of York, "Database of abstracts of reviews of effects (DARE)," 2019, 29 Nov, 2019. [Online]. https://www.crd.york.ac.uk/CRDWeb/AboutPage.asp

[4] D. Budgen, P. Brereton, S. Drummond, and N. Williams, "Reporting systematic reviews: Some lessons from a tertiary study," *Information and Software Technology*, Vol. 95, 2018, pp. 62–74.

[5] D. Costal, C. Farré, X. Franch, and C. Quer, "How tertiary studies perform quality assessment of secondary studies in software engineering," in *Proceedings of the XXIV Iberoamerican Conference on Software Engineering*. Curran Associates Inc., 2021, p. 1.

[6] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," School of Computer Science and Mathematics, Keele University, Keele, UK, Keele, UK, Tech. Rep., 2007.

[7] B.J. Shea, B.C. Reeves, G. Wells, M. Thuku, C. Hamel et al., "AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both," *BMJ*, Vol. 358, 2017.

[8] J.P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li et al., *Cochrane handbook for systematic reviews of interventions*. Wiley, January 2019.

[9] B.J. Shea, L.M. Bouter, J. Peterson, M. Boers, N. Andersson et al., "External validation of a measurement tool to assess systematic reviews (AMSTAR)," *PLoS One*, Vol. 2, No. 12, 2007, p. e1350.

[10] B.U. Burda, H.K. Holmer, and S.L. Norris, "Limitations of a measurement tool to assess systematic reviews (AMSTAR) and suggestions for improvement," *Systematic Reviews*, Vol. 5, No. 1, 2016, p. 58.

[11] U. Wegewitz, B. Weikert, A. Fishta, A. Jacobs, and D. Pieper, "Resuming the discussion of AMSTAR: What can (should) be made better?" *BMC Medical Research Methodology*,

Vol. 16, No. 1, Dec. 2016, pp. 111, s12 874–016–0183–6. [Online]. http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0183-6

[12] J.A.C. Sterne, J. Savović, M.J. Page, R.G. Elbers, N.S. Blencowe et al., "RoB 2: A revised tool for assessing risk of bias in randomised trials," *BMJ*, Vol. 366, 2019.

[13] T. Greenhalgh, "How to read a paper: Papers that summarise other papers (systematic reviews and meta-analyses)," *BMJ*, Vol. 315, No. 7109, 1997, pp. 672–675.

[14] K.S. Khan, G. Ter Riet, J. Glanville, A.J. Sowden, J. Kleijnen et al., "Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews," University of York, Tech. Rep. 4, 2001.

[15] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner et al., "Systematic literature reviews in software engineering – A tertiary study," *Information and Software Technology*, Vol. 52, No. 8, Aug. 2010, pp. 792–805.

[16] D.S. Cruzes and T. Dybå, "Research synthesis in software engineering: A tertiary study," *Information and Software Technology*, Vol. 53, No. 5, 2011, pp. 440–455.

[17] I. Nurdiani, J. Börstler, and S.A. Fricker, "The impacts of agile and lean practices on project constraints: A tertiary study," *Journal of Systems and Software*, Vol. 119, 2016, pp. 162–183.

[18] N.B. Ali, K. Petersen, and C. Wohlin, "A systematic literature review on the industrial use of software process simulation," *Journal of Systems and Software*, Vol. 97, 2014, pp. 65–85.

[19] T. Dybå and T. Dingsøyr, "Strength of evidence in systematic reviews in software engineering," in *Proceedings of the Second International Symposium on Empirical Software Engineering and Measurement, ESEM*, H.D. Rombach, S.G. Elbaum, and J. Münch, Eds. ACM, 2008, pp. 178–187.

[20] D.F. Stroup, J.A. Berlin, S.C. Morton, I. Olkin, G.D. Williamson et al., "Meta-analysis of observational studies in epidemiology: a proposal for reporting," *JAMA*, Vol. 283, No. 15, 2000, pp. 2008–2012.

[21] N.B. Ali and M. Usman, "A critical appraisal tool for systematic literature reviews in software engineering," *Information and Software Technology*, Vol. 112, 2019, pp. 48–50.

[22] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," *Information and Software Technology*, Vol. 106, 2019, pp. 201–230.

[23] B.A. Kitchenham, L. Madeyski, and D. Budgen, "SEGRESS: Software engineering guidelines for reporting secondary studies," *IEEE Transactions on Software Engineering*, 2022, p. 1.

[24] J.S. Molléri, K. Petersen, and E. Mendes, "CERSE-Catalog for empirical research in software engineering: A systematic mapping study," *Information and Software Technology*, Vol. 105, 2019, pp. 117–149.

[25] B.A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-based software engineering and systematic reviews*, Vol. 4. CRC Press, 2015.

[26] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, Vol. 64, 2015, pp. 1–18.

[27] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2014, pp. 38:1–38:10.

[28] B.A. Kitchenham, T. Dyba, and M. Jorgensen, "Evidence-based software engineering," in *Proceedings. 26th International Conference on Software Engineering*. IEEE, 2004, pp. 273–281.

[29] F.Q. Da Silva, A.L. Santos, S.C. Soares, A.C.C. França, and C.V. Monteiro, "A critical appraisal of systematic reviews in software engineering from the perspective of the research questions asked in the reviews," in *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. IEEE-ACM, 2010, pp. 1–4.

[30] M. Riaz, M. Sulayman, N. Salleh, and E. Mendes, "Experiences conducting systematic reviews from novices' perspective," in *14th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. BCS Learning & Development, Ltd., Swindon United Kingdom, 2010, pp. 1–10.

[31] E. Mendes, C. Wohlin, K.R. Felizardo, and M. Kalinowski, "When to update systematic literature reviews in software engineering," *Journal of Systems and Software*, Vol. 167, 2020, p. 110607.

[32] B.J. Shea, J.M. Grimshaw, G.A. Wells, M. Boers, N. Andersson et al., "Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews," *BMC Medical Research Methodology*, Vol. 7, No. 1, 2007, p. 10.

[33] A. Gates, M. Gates, G. Duarte, M. Cary, M. Becker et al., "Evaluation of the reliability, usability, and applicability of AMSTAR, AMSTAR 2, and ROBIS: Protocol for a descriptive analytic study," *Systematic Reviews*, Vol. 7, No. 1, 2018, p. 85.

[34] D. Pieper, R.B. Buechter, L. Li, B. Prediger, and M. Eikermann, "Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties," *Journal of Clinical Epidemiology*, Vol. 68, No. 5, 2015, pp. 574–583.

[35] C. Zapata, "Integration of usability and agile methodologies: A systematic review," *Design, User Experience, and Usability: Design Discourse*, 2015, pp. 368–378.

[36] M. Turner, B. Kitchenham, P. Brereton, S. Charters, and D. Budgen, "Does the technology acceptance model predict actual use? A systematic literature review," *Information and Software Technology*, Vol. 52, No. 5, 2010, pp. 463–479.

[37] O. Dieste and N. Juristo, "Systematic review and aggregation of empirical studies on elicitation techniques," *IEEE Transactions on Software Engineering*, Vol. 37, No. 2, 2010, pp. 283–304.

[38] A. Idri, F. Azzahra Amazal, and A. Abran, "Analogy-based software development effort estimation: A systematic mapping and review," *Information and Software Technology*, Vol. 58, 2015, pp. 206–230.

[39] M. Daneva and B. Lazarov, "Requirements for smart cities: Results from a systematic review of literature," in *12th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2018, pp. 1–6.

[40] D. Ameller, X. Burgués, O. Collell, D. Costal, X. Franch et al., "Development of service-oriented architectures using model-driven development: A mapping study," *Information and Software Technology*, Vol. 62, 2015, pp. 42–66.

[41] M.U. Khan, S. Sherin, M.Z. Iqbal, and R. Zahid, "Landscaping systematic mapping studies in software engineering: A tertiary study," *Journal of Systems and Software*, Vol. 149, 2019, pp. 396–436.

[42] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli, "Usability in agile software development: A tertiary study," *Computer Standards and Interfaces*, Vol. 64, 2019, pp. 61–77.

[43] H. Cadavid, V. Andrikopoulos, and P. Avgeriou, "Architecting systems of systems: A tertiary study," *Information and Software Technology*, Vol. 118, 2020, p. 106202.

[44] L. Yang, H. Zhang, H. Shen, X. Huang, X. Zhou et al., "Quality assessment in systematic literature reviews: A software engineering perspective," *Information and Software Technology*, 2020, p. 106397.

[45] C. Wohlin, "Writing for synthesis of evidence in empirical software engineering," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '14, ACM-IEEE. New York, NY, USA: Association for Computing Machinery, 2014, pp. 46:1–46:4.

[46] N.B. Ali, E. Engström, M. Taromirad, M.R. Mousavi, N.M. Minhas et al., "On the search for industry-relevant regression testing research," *Empirical Software Engineering*, Vol. 24, No. 4, 2019, pp. 2020–2055.

[47] N.B. Ali and K. Petersen, "Evaluating strategies for study selection in systematic literature studies," in *Proceedings of ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM*. ACM-IEEE, 2014, pp. 45:1–45:4.

[48] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, Vol. 55, No. 12, 2013, pp. 2049–2075.

[49] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell et al., *Experimentation in software engineering*. Springer Science and Business Media, 2012.

[50] M. Höst and P. Runeson, "Checklists for software engineering case study research," in *Proceedings of the First International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 2007, pp. 479–481.

[51] R.J. Wieringa, "Towards a unified checklist for empirical research in software engineering: first proposal," in *Proceedings of the 16th International Conference on Evaluation and Assessment in Software Engineering, EASE*. IET, 2012, pp. 161–165.

[52] N. Condori-Fernandez, R.J. Wieringa, M. Daneva, B. Mutschler, and O. Pastor, "An experimental evaluation of a unified checklist for designing and reporting empirical research in software engineering," Centre for Telematics and Information Technology (CTIT), The Netherlands, Tech. Rep., 2012.

Appendix A. QAISER instrument

Table A1: QAISER Instrument

1. Did the authors of the review adequately justify the need for undertaking the review? [22, 25, 26]		
For Partial Yes:	For Yes:	
The authors of the review should have:	As for Partial Yes, plus the authors of the review should also have ALL of the following:	
☐ Identified existing related reviews on the topic, or explained that no related review exists.	☐ Discussed related existing reviews on the topic, if any. ☐ Established a scientific or practical need for their review. [25]	☐ Yes ☐ Partial Yes ☐ No
Comments:		
2. Did the authors of the review establish a protocol prior to the conduct of the review? [7, 22, 25]		
For Partial Yes:	For Yes:	
The authors of the review confirm that a written protocol before the conduct of the review was established and is publicly available that provides details of the main elements of the systematic review process including the following:	As for Partial Yes, plus ALL of the following:	
☐ Appropriate review questions [7, 25, 26]. ☐ Search process [7, 25]. ☐ Study selection process [7, 22, 25]. ☐ Data extraction process [25]. ☐ Study quality assessment process (not relevant to most systematic mapping studies) [7, 25]. ☐ An outline of the data synthesis plan [7, 25].	☐ Either authors of the review report that there no deviations from the protocol or any deviations are documented and justified [7, 22, 25]. ☐ The protocol should have been internally validated by piloting selection criteria, search strings, data extraction and synthesis processes [25].	☐ Yes ☐ Partial Yes ☐ No
Comments:		

Table A1 continued

3. Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?		
For Yes: the review should have ALL of the following:		
☐ Reported the inclusion and exclusion criteria [3, 22, 25]. ☐ The criteria are aligned with the review questions. ☐ Provided appropriate justifications for any restrictions used in the inclusion and exclusion criteria (e.g., topic-related scoping restrictions, time-frame, language, study type, and peer reviewed works only) [7, 25]		. ☐ Yes ☐ No
Comments:		
4. Did the authors of the review use a comprehensive literature search strategy? [3, 7, 22, 25]		
When database search is used as the main method		
For Partial Yes:	For Yes:	
The review should have ALL of the following:	As for Partial Yes, plus the review should also have used:	
☐ An appropriate process for constructing the search strings including piloting [22, 25]. ☐ The search date is appropriate, i.e., the review includes sufficiently new papers in relation to the paper submission date. ☐ Search process validation based on using a known-set of papers [22, 25, 26] and providing an insightful discussion concerning why some papers are missing. ☐ At least one relevant indexing database (e.g., Scopus) in combination with relevant publisher databases (e.g., IEEE and ACM) [25]. ☐ Appropriately documented the search process (e.g., known-set, search strings, and search results) [2, 22, 25].	☐ At least one additional search method (e.g., snowballing, manual search, or use DBLP or Google Scholar of key researchers) [7, 22, 25].	☐ Yes ☐ Partial Yes ☐ No
When snowballing is used as the main method		
For Partial Yes:	For Yes:	

Table A1 continued

The review should have ALL of the following:	As for Partial Yes, plus the review should also have used at least ONE of the following:	
☐ Appropriately justified the use of snowballing as the main method [25]. ☐ Selected an appropriate start/seed set. The selection process should be explained and justified [27]. ☐ Search process validation based on using a known-set of papers [22, 25, 26] and providing an insightful discussion concerning why some papers are missing. ☐ Performed an acceptable number of backward and forward snowballing iterations [27]. ☐ Appropriately documented the search process (e.g., start/seed set, known-set, and search results) [2, 22, 25].	☐ At least one additional search method (e.g., manual search, or use DBLP or Google Scholar of key researchers) [25, 27]. ☐ Snowballin g iterations until no new papers were found [27].	☐ Yes ☐ Partial Yes ☐ No
Comments:		
5. Did the authors of the review use a reliable study selection process? [7, 25, 26]		
For Partial Yes:	For Yes:	
☐ At least two authors of the review selected a representative sample of eligible studies, achieved good agreement, and reported the agreement level, with the remainder selected by one review author [7, 22, 25].	☐ At least two authors of the review independently agreed on selection of eligible studies and reached consensus on which studies to include [7, 25, 26]. OR ☐ As with Partial Yes, with the additional requirement that at least two authors reviewed and achieved consensus about all excluded studies.	☐ Yes ☐ Partial Yes ☐ No
Comments:		
6. Did the authors of the review use a reliable data extraction process? [7, 22, 25, 26]		.
For Partial Yes:	For Yes:	

Table A1 continued

☐ At least two authors of the review extracted data from a sample of included studies, achieved good agreement, and reported the agreement level, with the remainder extracted by one review author [7, 25, 26].	☐ At least two authors of the review achieved consensus on which data to extract from the included studies [7, 22, 25, 26].	☐ Yes ☐ Partial Yes ☐ No
Comments:		

7. Did the authors of the review discuss and justify the exclusion of the potentially relevant studies that were read in full text? [7, 25]

For Partial Yes:	For Yes:	
The review should have either ONE of the following:	As for Partial Yes, plus the review should also have the following:	
☐ Provided a list of all potentially relevant studies that were read in full text, but excluded from the review [7, 25].	☐ Justified the exclusion from the review of each potentially relevant study that was read in full text [7].	☐ Yes ☐ Partial Yes ☐ No
Comments:		

8. Did the authors of the review provide sufficient primary studies' characteristics to interpret the results? [3, 7, 22, 25]

For Yes, the review should have described ALL of the following:	
☐ Populations [7] ☐ Interventions, when applicable [7] ☐ Outcomes, when applicable [7] ☐ Study types [7] ☐ Study contexts [7]	☐ Yes ☐ No
Comments:	

9. Did the authors of the review use an appropriate instrument for assessing the quality of primary studies that were included in the review? [7, 25]

For Yes, the review should have used appropriate instruments for different types of studies included in the review. An appropriate instrument would have questions related to ALL of the following [25]:

Table A1 continued

☐ *"The goals, research questions, hypotheses and outcome measures"* [25]. ☐ *"The study design and the extent to which it is appropriate to the study type"* [25]. ☐ *"Study data collection and analysis and the extent to which they are appropriate given the study design"* [25]. ☐ *"Study findings, the strength of evidence supporting those findings, the extent to which the findings answer the research questions, and their value to practitioners and researchers"* [25]	. ☐ Yes ☐ No
Comments:	

10. Did the authors of the review use a reliable quality assessment process? [25]

For Partial Yes:	For Yes:	
☐ At least two authors of the review performed quality assessment of a representative sample of eligible studies and achieved good agreement, with the remainder performed by one review author [25].	☐ At least two authors of the review independently performed quality assessment of eligible studies and reached consensus about the quality levels or scores of the eligible studies[25].	☐ Yes ☐ Partial Yes ☐ No
Comments:		

11. Were the primary studies appropriately synthesized? [3, 4, 22, 25]

For Yes, the review should have ALL of the following:	
☐ Selected an appropriate synthesis method given the review questions and extracted data [4, 22, 25]. ☐ Applied the selected synthesis method appropriately. ☐ Provided a clear trace linking the answers of review questions and interpretations to the data from the primary studies.	☐ Yes ☐ No
Comments:	

12. Did the authors of the review investigate the impact of the quality of individual studies on the results of the review? [7, 22, 25]

For Yes, either ONE of the following:	
☐ Included only high-quality studies [7, 25]. OR ☐ The authors have discussed the impact of differences in the quality of individual studies on the results of the review [7, 25].	☐ Yes ☐ No

Table A1 continued

Comments:		
13. Did the authors of the review investigate the impact of primary studies' characteristics on the results of the review?		
For Yes, either ONE of the following:		
☐ There were no significant similarities or differences to warrant a separate analysis. OR ☐ The authors have discussed the impact of primary studies' characteristics on the results of the review [7, 25].		☐ Yes ☐ No
Comments:		
14. Did the authors of the review provide appropriate recommendations and conclusions from the review? [4]		
For Partial Yes:	For Yes:	
The review should have the following:	As for Partial Yes, plus the recommendations and conclusions should also be:	
☐ Provided satisfactory recommendations and conclusions based on the review results.	☐ Clearly traceable back to the review results. ☐ Clearly targeting specific stakeholders. ☐ Well aligned with the upfront motivation for undertaking the review, or are any deviations well explained. ☐ Providing new valuable insights to the community.	☐ Yes ☐ Partial Yes ☐ No
Comments:		
15. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review? [7, 25]		
For Yes, either ONE of the following:		
☐ The authors reported no competing interests. OR ☐ The authors described their funding sources [7, 25] and how they managed any other potential conflicts of interest (e.g., how/who handled their own publications on the review topic while conducting the review) [7].		☐ Yes ☐ No
Comments:		

Appendix B. QAISER guidance document

In this document, we provide further guidance to support a consistent interpretation of items in QAISER.

Item 1: Did the authors of the review adequately justify the need for undertaking the review?

A large number of SLRs are reported in software engineering every year. A review should be initiated on the basis of a practical or scientific need. The authors of the review should also extensively search for any existing reviews or mapping studies on the topic. The authors should only continue planning the review if there are no existing ones that are up to date on the specific area [25]. Mendes et al. [31] provide support to decide if a review should be updated.

To score "Partial Yes", appraisers should check if the authors of the review have made a sufficiently extensive effort to identify related reviews on the topic. For example, a search using keywords for systematic secondary studies (like "systematic review", "systematic literature review", "systematic mapping study", or "systematic map") and topic specific keywords. For examples of such search please consult [18, 46].

To score "Yes", in addition to the criterion under "Partial Yes", appraisers should ensure that the authors of the review have established the need for undertaking the review. If there are existing reviews on the same topic, the authors need to establish the need by highlighting the gap in the existing reviews, and explaining how their review is going to fill the gap. In case there are no existing reviews on the topic, the authors explain why is it essential to aggregate the evidence on the topic.

The information about the need for review is typically described in the background or related work sections of the report.

Item 2: Did the authors of the review establish a protocol prior to the conduct of the review?

To reduce the risk of bias, it is important that the authors of the review have developed and validated a written protocol before commencing the review.

To score "Partial Yes", appraisers should first ensure that the authors of the review confirmed the establishment of a protocol before the conduct of the review. In addition, the protocol should be either accessible online and the review report describes where and how can it be accessed, or important parts of the protocol are included in the review report. Furthermore, the protocol should have documented appropriate review questions, the processes for search, study selection, data extraction, quality assessment and at least an outline for the data synthesis plan.

To rate "Yes", appraisers should first check that all of the criteria under "Partial Yes" have been met. In addition, the authors of the review should have clearly documented and justified any deviations from the protocol and discuss their impact on the study. Lastly, appraisers should also ensure that the protocol has been validated internally by pilot testing of different review processes (e.g., trial searches, selection, data extraction).

The above information about the protocol is typically described in the methodology section of the review report.

Item 3: Did authors of the review report their inclusion and exclusion criteria and, explain and justify them in terms of the review questions?

A review should use documented selection criteria [3, 7, 22, 25, 26].

To score "Yes", appraisers should ensure that the authors of the review have justified any restrictions, e.g., on research designs, the time frame of publication, and the type of

publications imposed in the selection process. Furthermore, the justification should also address the likely impact of the restrictions on the studied population and the generalization of the findings.

The selection criteria and the justifications for any restrictions are expected to be found in the methodology or limitations/threats to the validity section of the review report. Furthermore, some of the exclusion criteria may have been implemented in the search process.

Item 4: Did the authors of the review use a comprehensive literature search strategy?

A comprehensive search strategy is important to maximize the coverage of the relevant literature. The authors of the review should provide a justification for using a particular search method (e.g., database or indexing service search or snowballing) as a primary method for searching the relevant literature.

To rate "Partial Yes", appraisers should check the following in case of a database or indexing service search as the primary search method:

- The authors of the review have used an appropriate process for identifying the search terms, synonyms and constructing the search strings.
- The authors have included sufficiently new papers given the paper submission date. AMSTAR 2 recommends that the last search was conducted within the last 24 months. However, a different time may be more appropriate for the area of the review being assessed. The time may, for example, depend on the speed of the evolution of a specific area of research.
- The authors of the review have validated their search process by comparing their search results with a known-set of papers. The known-set of papers are the relevant papers that are already known to the authors of the review based on, for example, manual search or their knowledge of the review topic. The validation should, if papers are missing, include an insightful discussion concerning why some papers are missing.
- The authors of the review have used a combination of publisher databases and indexing services. IEEE and ACM are the most relevant publisher databases in software engineering, as they publish the most important and relevant conferences and journals in software engineering [25]. As a minimum, the authors should have used IEEE and ACM among the publisher databases and one indexing service (e.g., Scopus). Lastly, an important aspect of the application of the search strings to the selected databases is the timeliness of the search process. Appraisers should also ensure that the reported search is not outdated. While making this assessment of the timeliness of the search process, appraisers need to account for the time required to complete the remaining steps in the review process, writing the review report, and the peer-review process.
- The authors of the review have documented the search process. Appropriate documentation of the search process is important to ensure repeatability and transparency. The authors of the review should document: general and database specific search strings, total and database specific search results, search filters (e.g., years) used, date when the search strings were applied, known-set of papers used for validation, and validation measures (i.e., recall and precision) for details see [2].

To rate "Yes", in addition to the above criteria, the authors of the review should have also used at least one additional search method (e.g., snowballing).

To rate "Partial Yes", appraisers should check the following in case of a snowballing search as the primary search method:

- The authors of the review have used an appropriate process for identifying the seed set for starting the snowballing procedure. The way of identifying the seed set is well-documented and -motivated.
- The authors have included sufficiently new papers given the paper submission date. AMSTAR 2 recommends that the last search was conducted within the last 24 months. However, a different time may be more appropriate for the area of the review being assessed. The time may, for example, depend on the speed of the evolution of a specific area of research.
- The authors of the review have validated their search process by comparing their search results with a known-set of papers. The known-set of papers are the relevant papers that are already known to the authors of the review based on, for example, manual search or their knowledge of the review topic. The validation is performed by computing recall and precision using the search results and the known set of relevant papers (for details, refer to [25]).
- The authors have iterated the snowballing procedure until no more papers are found.
- The authors of the review have documented the search process. Appropriate documentation of the search process is important to ensure repeatability and transparency. The authors of the review should document: identification of the seed set, the different iterations conducted, known-set of papers used for validation, and validation measures (i.e., recall and precision).

To rate "Yes", in addition to the above criteria, the authors of the review should have also used at least one additional search method (e.g., manual search of key journals or conference proceedings, or use DBLP or Google Scholar of key researchers) or continuing snowballing iterations until no new papers are found).

Item 5: Did the authors of the review use a reliable study selection process?
To reduce bias and the possibility of making mistakes, the crucial step of inclusion and exclusion of the papers should involve at least two reviewers [7, 47].

To rate "Partial Yes", appraisers should check if at least two authors of the review selected a representative sample of eligible studies, achieved good agreement, and also reported the agreement level. While reporting the agreement level, the review report should also describe the sample size and the process adopted for achieving a good agreement level. Only after a Kappa score indicates that a good agreement has been achieved between at least two authors on a representative sample, a single author can proceed with the selection process for the remaining studies.

To rate "Yes", appraisers should check that one of the two following processes are followed during study selection: 1) two authors of the review independently performed study selection on all eligible studies and reached consensus on which studies to include or exclude, 2) as with Partial Yes, i.e., two authors of the review selected a sample of eligible studies, achieved and reported good agreement level, with the remainder selected by one review author, but in that case all excluded studies must be reviewed by at least two authors of the review. A single reviewer should only proceed with the selection after a Kappa score indicating strong agreement between multiple authors of the review has been reached. However, even in this case, the excluded studies should be reviewed by at least one more author of the review – to ensure that decision to exclude any study is not made by a single author. The review would suffer from threats to validity if some studies were excluded by a single researcher. when it comes to including potentially irrelevant studies by a single author, the risk is mitigated by the fact that the irrelevance of such studies will become visible during the data extraction and synthesis processes. Therefore, it is important

to ensure that all excluded studies are reviewed by at least two authors. Appraisers should also check that the rules for inclusion and exclusion, and how these rules were applied and how any differences between reviewers were resolved are described. Furthermore, the report should also report the number of papers remaining at each stage [4].

The information about the study selection process is expected to be described in the methodology and results sections of the review.

Item 6: Did the authors of the review use a reliable data extraction process?

To ensure repeatability of the study and to avoid bias, it is important that the data extraction is not solely performed by a single researcher.

To rate "Partial Yes", appraisers should check at least two authors extracted data from a sample of studies and have achieved a good agreement. The report should also include the agreement level level achieved, size of the sample, and the process used to achieve consensus on which data to extract.

To rate "Yes", appraisers should ensure that data is extracted by at least two authors of the review. It is important to check that the review report provides a description of the process used to achieve consensus and shared understanding on which data to extract.

For SLRs that use qualitative synthesis methods, the data extraction process should be designed to minimize dependence on the viewpoint of a single member of the systematic review team and maximize the opportunities for team-based working and team decision making. Any tools used to assist data extraction and analysis should be identified and their contribution to data extraction process explained. The information of the data extraction process is generally described in the methodology section of the review report.

Item 7: Did the authors of the review discuss and justify the exclusion of the potentially relevant studies that were read in full text?

This item refers to studies that were deemed relevant by authors of the review on a reading of the title and abstracts. However, after full-text reading, the authors concluded that the papers are not relevant to the current review. It is expected that the authors of the review should document such papers along with the reason for their exclusion. This will help increase confidence in the results, allow reflecting on the selection criteria used in the study, allow replications, and enable further research (for example, by leveraging on the filtered list of papers for a different analysis).

To rate "Partial Yes", appraisers should see that the authors of the review have provided a list of such potentially relevant papers. Alternatively, the authors of the view should have reflected on the main reasons (e.g., papers that report no data for any of the review questions, papers excluded due to low quality) for excluding the papers that were read in full text.

In order to rate "Yes", in addition to one of the alternatives for "Partial Yes", justifications for excluding the potentially relevant papers should also be provided.

This documentation (i.e., a list of potentially relevant papers that were excluded after full-text reading and justifications for excluding them) can be made available in an appendix or as supplementary material for review online (along with other supporting material like the review's protocol).

Item 8: Did the authors of the review provide sufficient primary studies' characteristics to interpret the results?

The relevance and reliability of a systematic review depends, besides other factors, also on a number of factors related to the included studies such as its type (e.g., case study, survey, and experiment), context (real life or laboratory setting), participants (practitioners or students), and publication venue (e.g., a reputable conference/journal). The review report should describe adequate details about the characteristics of the included studies

to inform the review readers about the kind of evidence that is used to draw conclusions. The concept of population in SE empirical studies is not limited to human subjects. In SE empirical studies, the focus may be on other items of interest – e.g., artifacts, issues or other events. Therefore, depending on the context, appraisers need to see what is a relevant population and whether or not authors of the review have included enough details about it in their review report.

To rate "Yes", appraisers should ensure that the authors of the review have provided enough details about the population, interventions (when relevant), outcomes (when relevant), research designs and settings of the included studies.

These details may not be described at one place in a review report, and therefore could be challenging to find. Normally, part of this information is described in the start of the results section in a review report.

Item 9: Did the authors of the review use an appropriate instrument for assessing the quality of primary studies that were included in the review?

Due to several reasons, including the variety of research designs used in primary studies, reporting quality, use of inconsistent terminology, etc., quality assessment is a challenging task in software engineering systematic literature reviews [25, 48]. Several research-design specific checklists (e.g., for experimentation [49] and case study research [50]) and generic instruments (e.g., [51, 52]) have been proposed in literature. However, as concluded by Kitchenham [25], it is not feasible to use the same instrument to assess the quality of different types of studies.

To rate "Yes", appraiser should ensure that the choice of the instruments used (whether an existing one or one formulated by the authors of the review) has been justified given the goals of the SLR and nature of included studies. Furthermore, the instrument used is expected to evaluate at least the research design, data collection, analysis reporting, and the strength of evidence given the stated goals of the primary study.

The information on the quality assessment of the primary studies is expected to be described in the methodology and results sections of the review report.

Item 10: Did the authors of the review use a reliable quality assessment process?

Like Items 5 and 6, it is important that the quality assessment is not performed solely by a single author of the review.

To rate "Partial Yes", appraisers should check if at least two authors have performed pilot quality assessment of a sample of the included studies to evaluate the objectivity of the quality assessment criteria and to develop a shared understanding of it.

To rate "Yes", appraisers should see that at least two authors of the review independently performed the quality assessment of either all included studies or a sample of included studies (with the remaining performed by one review author) and achieved good consensus. The review report should also describe how differences were resolved in case of different quality scores.

The information about the quality assessment process is typically expected in the methodology section of the review report.

Item 11: Were the primary studies appropriately synthesized?

Synthesis is one of the most important and also challenging parts of a systematic literature review. Without synthesis, the review would be of limited use.

In order to score "Yes", appraisers need to see if the authors of the review have used and justified an appropriate method for synthesis. It may be the case that the authors of the review do not use the correct or appropriate name for the used synthesis method [16]. In that case, appraisers would have to carefully read the review report in order to

make a decision on this item. The appraisers should further check if the selected synthesis method was appropriately applied and that there is a clear chain of evidence from the answers to the research questions to the data from the primary studies.

The information about the synthesis method and its output may be documented in a separate section. In some cases it may be described in the discussion section after the results section. It could also be the case that the justification for selecting a specific synthesis method is described in the research methodology section, while the outputs of the synthesis step are described in a separate section.

Item 12: Did the authors of the review investigate the impact of the quality of individual studies on the results of the review?

A review should take the quality of the individual studies into account when interpreting the results. This will increase the confidence in the findings and conclusions of the review.

To rate "Yes", appraisers should see that either the review has excluded studies that do not meet the quality criteria defined in the study, or the authors have investigated and discussed the impact of differences in the quality of included studies on the results of the review. In the case of the first option, i.e., if the authors have only included high quality studies, the differences in the quality of study are not expected to be as high as requiring a further impact analysis. In the case of the second option – i.e., if the low quality studies are not excluded by the authors of the review, it is important that the authors investigate the impact of the quality of included studies on the results of the studies' synthesis by – for example – categorizing the results and analysis based on the quality of included studies. The information on using the quality of studies while interpreting the results is expected to be described in the discussion or analysis sections of the review report where results are further discussed/analyzed to draw conclusions.

Item 13: Did the authors of the review investigate the impact of primary studies' characteristics on the results of the review?

There are many factors that can cause heterogeneity in the results of the included studies. It is important to analyze the causes of the heterogeneity in results, if any, while interpreting the results and drawing any conclusions. For example, it could be the variations in the contextual factors (e.g., student versus practitioners as subjects) that lead to differences in the results of different studies. Furthermore, quality scores or some specific quality criteria might also help in explaining the heterogeneity observed in the results [25]. This item is concerned with the use of study characteristics in Item 8.

In order to rate "Yes", appraisers should see that the authors of the review have investigated the impact of study characteristics (including the number of studies) on the results of the review.

This discussion is likely to be found after the results section of the review report.

Item 14: Did the authors of the review provide appropriate recommendations and conclusions from the review?

The usefulness of results of the review for the target stakeholders is critical to assess the relevance of the review. This item is a reflection on the aims as motivation for the review assessed in the first item of the instrument (i.e., item 1).

For "Partial Yes" the review should have satisfactory recommendations and conclusions based on the review results.

For "Yes", in addition providing satisfactory recommendations and conclusions, the recommendations and conclusions from the review shall also be traceable to the review results, clearly targeting specific stakeholders, well aligned with the motivation and provides new insights to the community.

Item 15: Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?
To ensure the reliability of a review, it is important that the authors of the review report their sources of funding and any other conflicts of interest. The disclosure of the sources of funding is quite obvious. However, identifying other types of conflicts of interest is not that straightforward.

For example, if the authors of the review have published on the topic of the review or have a vested interest in the outcome of the review, there is a potential for bias when selecting, analyzing and interpreting their own work and studies with competing alternatives.

It is encouraged that authors of the review should be experts in the topic area, so it is common that they have published extensively in the topic area. Thus, it is important that the authors of the review report their effort in identifying any conflicts of interest they have, which is relevant for the review. A mitigation strategy in this case is to establish a process and have it reviewed by independent researchers not participating in the literature review.

To rate "Yes", the appraisers should ensure that the authors of the review have reported on the presence or absence of any conflicts of interest. In case there was some conflict of interest, the authors of the review should have described and justified the steps taken to mitigate the threat of bias in the results of the review.

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230106, DOI: 10.37190/e-Inf230106

Value-based Software Engineering: A Systematic Mapping Study

Norsaremah Salleh*, Emilia Mendes**, Fabiana Mendes***, Charitha Dissanayake Lekamlage****, Kai Petersen****

*Department of Computer Science, Kulliyyah of ICT, International Islamic University Malaysia
**Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden
***Faculty UnB Gama, University of Brasilia, Brasilia-DF, Brazil
****Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden

norsaremah@iium.edu.my, emilia@bth.se, fabianamendes@unb.br, dilukshidissanayake@yahoo.com, kai.petersen@bth.se

Abstract

Background: Integrating value-oriented perspectives into the principles and practices of software engineering is fundamental to ensure that software development activities address key stakeholders' views and also balance short-and long-term goals. This is put forward in the discipline of value-based software engineering (VBSE)
Aim: This study aims to provide an overview of VBSE with respect to the research efforts that have been put into VBSE.
Method: We conducted a systematic mapping study to classify evidence on value definitions, studies' quality, VBSE principles and practices, research topics, methods, types, contribution facets, and publication venues.
Results: From 143 studies we found that the term "value" has not been clearly defined in many studies. VB Requirements Engineering and VB Planning and Control were the two principles mostly investigated, whereas VB Risk Management and VB People Management were the least researched. Most studies showed very good reporting and relevance quality, acceptable credibility, but poor in rigour. Main research topic was Software Requirements and case study research was the method used the most. The majority of studies contribute towards methods and processes, while very few studies have proposed metrics and tools.
Conclusion: We highlighted the research gaps and implications for research and practice to support VBSE.

Keywords: Systematic mapping, value-based software engineering, VBSE

1. Introduction

Value-based Software Engineering (VBSE) aims to incorporate value thinking into the wide range of Software Engineering principles and practices [1]. It opposes a value-neutral approach to SE practice and research, where value-neutral is described as [1]:

– "Every requirement, use case, object, and defect is treated as equally important";
– "Methods are presented and practiced as largely logical activities involving mappings and transformations (e.g., object-oriented development)";

Submitted: 09 Mar. 2023; Revised: 21 May 2023; Accepted: 30 May 2023; Available online: 12 Jun. 2023

– "'Earned value' systems track project cost and schedule, not stakeholder or business value";
– "A 'separation of concerns' is practiced, in which the responsibility of software engineers was 'confined to turning software requirements into verified code', rather than to continuously maintain the consistency along the chain of evolving value propositions, system and software requirements, architecture and code."

Furthermore, one of main criticisms towards a value-neutral view is that it can also deteriorate projects' outcomes [1, 2]. A value perspective should be integrated into the full range of existing and emerging SE principles and practices, such as value-based requirements engineering, architecting, design and development, verification and validation, planning and control, risk management, quality management, and people management [1]. Finally, VBSE should be the basis for a framework in which the previously mentioned SE principles and practices "compatibly reinforce each other" [1, 2]. It is important to note that the context of "value" in this study refers to the broader definition of value as used in [2], that define value as "relative worth, utility, or importance" [2], in addition to the traditional and common definition of value, i.e., in terms of economics or monetary aspects [3].

After Boehm's seminal paper's publication, other VBSE publications followed, investigating value-based approaches and techniques in SE such as in [4–6]. It is important to know to what extent the proposed approaches and techniques contribute to software development or are used by practitioners, and whether the interest in value-based studies still persists or not. Therefore, the goal and main research contribution of this paper is to detail a mapping study aimed to identify primary studies in VBSE. The motivation to conduct this mapping study is to understand the research efforts that have been put into VBSE by providing a catalogue or classification of evidence of VBSE research. These include understanding the definition and context of value used in the studies, their quality and rigour, the VBSE principles and practices studied, the research topics and the publication venue. Our mapping study structures the VBSE body of knowledge through a systematic classification of evidence based on the VBSE definition and agenda given by Boehm (2003) [1].

This mapping study's key contributions are to: i) analyze how value is defined in VBSE studies and the quality of those studies, measured according to four categories (reporting, rigor, credibility, and relevance); ii) identify and summarize trends in the VBSE research (related to SE principles and practices) and the research gaps for future research; iii) identify and summarize the main topics researched in the studies, and the research gaps and topics for future interest, looking at publication trends over time; iv) reveal gaps for future research concerning the use of research methods, maturity of research based on the type of investigation, and possible opportunities for research and contribution types; and v) present the publication venue. We also document important research gaps to better inform both practice and future research in this field. The research questions for this mapping study and the motivations for each question are outlined in Table 1.

The remainder of this paper is organized as follows: Section 2 presents the background of research and the related work. Section 3 describes the research method. Section 4 presents the results from the mapping study followed by a discussion and threats to validity in Section 5. Finally, Section 6 concludes our work.

Table 1. Research questions and motivation

RQ#	Research Question	Motivation(s)
RQ 1	How value has been defined in the existing VBSE studies?	Understanding how value is defined is central to know how value can be used or has been practiced in any levels of decision-making in SE.
RQ 2	What do we know about the quality of VBSE studies, particularly on the quality of reporting, rigor, credibility and relevance?	To measure the quality of VBSE studies using a well-known classification proposed by [7]. Researchers can use such information to focus follow-up systematic reviews on studies of high quality.
RQ 3	What are the SE principles and practices investigated so far in VBSE, and how has this changed over time?	Researchers and practitioners can identify relevant practices in their areas of interest (e.g., requirements and VBSE) based on the catalogue/classification concerning SE principles and practices.
RQ 4	What are the most investigated research topics in VBSE, and how has this changed over time?	Researchers and practitioners can identify relevant papers for specific research topics based on the catalogue/classification concerning topics.
RQ 5	What are the research methods used in VBSE studies and how many studies looked at each method (e.g., case study, experiments, survey, etc.)?	To reveal gaps for future research concerning the use of research methods (e.g., showing the needs for more industrial studies – e.g., case studies) in VBSE areas.
RQ 6	What are the research types that these studies apply (e.g., validation/evaluation/solution proposal, etc.) and how many studies looked at each research type?	To reveal gaps for future research concerning the types of research documented.
RQ 7	What contribution facets do they provide (e.g., process, method, model)?	To reveal gaps for future research concerning the contribution facets, (e.g., showing which contribution facet is lacking).
RQ 8	What are the publication venues for VBSE research?	To provide awareness about where VBSE papers have been published.

2. Background and related work

2.1. Concepts of VBSE

The value-based paradigm in SE has emerged when several authors promoted "value" as a basis for decision-making in software engineering rather than relying on "cost" alone (e.g., [8]. One of the arguments is that "value-neutral "approaches in software development are unable to deal with most of the sources of software project failure [9]. Under the "value-neutral" setting, the focus is more on the technical aspects such as quality, cost, and development time, and where decisions made about a software project are "de-coupled from the value propositions that established the project" [1]. Conversely, under a "value" setting, all participating stakeholders (e.g., customers, developers, managers, finance, marketing) must understand and handle each other's value propositions. Therefore, the goal is to create a product or service that adds value to all the stakeholders [1]. Hence, VBSE aims to bring value considerations more prominent so that software engineering decisions at all levels can be optimized to meet the objectives of the stakeholders [2].

Boehm [9] defined VBSE as "the explicit concern with value concerns in the application of science and mathematics by which the properties of computer software are made useful to people". The application of science includes both social and physical sciences, whereas

the mathematics perspectives include the utility theory, game theory, statistical decision theory, real options theory as well as logic, complexity theory, and category theory [9]. Since software is expected to be "made useful to people", the inclusion of economics, management science, cognitive sciences, and humanity are required to create a successful software system [2]. As such, VBSE is emphasized as "a multifaceted, multidisciplinary approach that covers all practices, activities, and phases involved in software development, addressing a wide variety of decisions about technical issues, business models, software development processes, software products and services, and related management practices." [2].

To address such multifaceted and multidisciplinary aspects of VBSE, an initial "4 + 1" theory of VBSE has been developed by Boehm and Jain [10]. The core of the theory is the stakeholder win-win Theory W (also known as the Enterprise Success Theorem), which states, "Your enterprise will succeed if and only if it makes winners of your success-critical stakeholders" [10]. The theory provides a process framework for guiding VBSE activities, addressing two major questions: "which values are important?" and "how is success measured?". The theory is supported by four other theories known as utility theory, decision theory, dependency theory, and control theory [10].

2.2. VBSE principles and practices

The aim of VBSE as a discipline is to integrate value-oriented perspective into all of the software engineering aspects such as requirements engineering, architecting, design and development, verification and validation, planning and control, risk management, quality management, and people management [1]. Hence, we used as basis the existing and emerging SE principles and practices outlined in the VBSE agenda [1] as follows:

- **Value-based requirements engineering:** Principles and practices to identify a system's success-critical stakeholders and to elicit and reconcile value propositions with respect to the system.
- **Value-based architecting:** Reconciliation of the system objectives with achievable architectural solutions.
- **Value-based design and development:** Techniques to ensure that software design and development process incorporates value considerations.
- **Value-based verification and validation (V&V):** Techniques to ensure a software solution satisfies its value objectives and provide ways to prioritize V&V tasks.
- **Value-based planning and control:** Incorporates the value delivered to stakeholders in terms of cost, schedule and product planning and control techniques.
- **Value-based risk management:** Incorporates value in identifying, analyzing, prioritizing and mitigating risk.
- **Value-based quality management:** Prioritizes desired quality factors that relate to stakeholders' value propositions.
- **Value-based people management:** Build stakeholder's team, manages expectation, reconciles stakeholder's value propositions, and integrates ethical considerations in a project's execution.
- **Theory of VBSE:** Application and development of theories in VBSE.

2.3. Related literature reviews

One of the earlier publications on VBSE was published as an edited book [2]. This book consists of fully refereed chapters providing foundations of VBSE, and mainly focusing

on software engineering decisions and their consequences from a value-based perspective. These include a presentation of state-of-the-art methods and techniques for evaluation of software products, services, processes and projects from an economic point of view. Additionally, the benefits of VBSE are also demonstrated through examples and case studies. This book, however, cannot be considered as a secondary study of VBSE research, but simply a compilation of chapters relating to a wide range of VBSE topics.

With regard to secondary studies, the first fully refereed publication we are aware of is by Khurum et al. [11] who performed a mapping study that relates to value but which sole focus was to identify value propositions or factors that have been used and should be considered while making decisions about software product development and management relating to software intensive products. They also included primary studies outside VBSE domain that were published in fields such as economics and marketing. The results from their mapping study were used to build a classification of value propositions called the Software Value Map (SVM). Their mapping study covered a period from 1969 to 2010, and has no overlap with ours. While our mapping study aims to provide a detailed overview of the VBSE domain, their study focused solely on identifying value propositions to be used to build the SVM classification.

The second fully-refereed secondary study we are aware of is Khan and Khan [12]. They presented a literature review focusing upon the impact that the adoption of a "value-based" approach to SE had upon software reusability and quality. Their analysis, limited to only ten (10) studies, presents different value-based pricing criteria, selection of automated tools, existing Component off the shelf (COTS), and conflict resolutions among different stakeholder value propositions. They reference Barry Boehm's work; however, some of the primary studies included in the paper are not "value-driven"; rather, they see value solely as financial, monetary value. This study also does not provide an overview of the VBSE domain, and, like [11], includes studies that are outside the VBSE domain.

The third study relating to this work is a much shorter version of this mapping study [13], which is a paper we published in 2019, covering three of the eight RQs described herein. The differences between what was reported in [13] and what is reported herein are as follows:

- The study in [13] presented results for the RQs 3 to 5, covering the period 2003 to 2017, and without investigating the change in trends over time. Herein we also present the results for RQs 3 to 5, however covering a wider period (2003 to 2020), and also investigating whether trends have changed over time (RQs 3 and 4).
- Herein we have also extended the number of RQs to 8, and covering a range of different aspects relating to VBSE research (e.g., value definition, studies quality, research types used, and contribution facets employed). We also include publication venues of VBSE research.

Finally, there is also a grey literature Masters thesis [14] that provides a mapping study of value in SE, by means of 364 studies published within the timeframe between 1990 and 2010. Although their study aimed to classify the contributions of VBSE studies and investigated the practical application and validation of the solutions in industry, quite a number of their included studies are non-VBSE studies. Further, the overall mapping focused on the software development process areas, the research types and contributions of the study. In our mapping study, we included data covering the VBSE agenda, value definition, the research quality, methods and publication venue of VBSE studies in addition to the contributions facets and the research types.

In summary, we believe that the additional set of RQs, and the extensions to the three RQs detailed in [13], provide a significant additional research contribution, when compared to what was documented in our previous work [13]. Furthermore, given that to date there is no fully-refereed and/or rigorously conducted mapping study that provides a detailed overview of VBSE, we see this as an additional and significant contribution of this work.

3. Research method

In conducting this mapping study, we refer to the guidelines presented in [15] and [16]. The activities involved are illustrated in Figure 1, which consist of three phases: planning, executing and reporting. Beside each activity, there is also the identification of each of the author(s) – 1 for first author, 2 to second author and so forth, who participated in that activity.

The planning phase relates to making decisions such as identifying the mapping study's goals, defining the scopes, research questions, search strategy, selection criteria, and defining

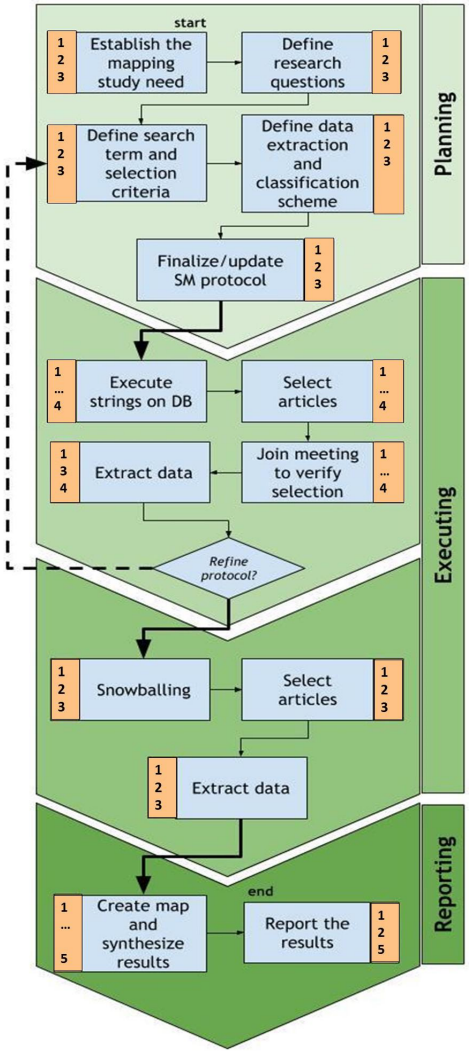

Figure 1. Systematic mapping process

the data extraction and classification process. In the second phase (Executing), the tasks include all the processes that relate to the mapping study's execution, which include study selection, and data extraction. All the first four authors have been involved in the selection of articles, and numerous joint meetings were held to verify study selection. The search results retrieved from online databases were entered to Parsifal (web-based tool). Once we finalized the list of included studies, a backward snowballing search was conducted. For validation of data extraction, we identified the disagreements in the extracted data and resolved through a joint meeting. The last phase – Reporting, represents the reporting and the results evaluation. The first, second, third and fifth authors were involved in this phase to synthesize and write-up the results. We carried out checks and balances through joint meetings held between the authors. We refer to the SEGRESS Guidelines [17] for reporting this mapping study. Our replication package is also available at the following link: https://zenodo.org/record/7901667#.ZF84uexByu5

3.1. Search strategy

Since our mapping study aims to search for relevant studies reporting value-based Software Engineering research, we used the following string:

```
((``value-based'' AND ``software engineering'') OR (``value-based software
    engineering'') OR (``value based'' AND ``software engineering'') OR
   (``Value based software engineering'') OR VBSE) OR ((value OR ``value
 based'' OR ``valuation'' OR ``value creation'') AND (``economics based''
     OR ``decision making'' OR economics OR ``software project'') AND
    (``software engineering'' OR software OR ``software development''))
```

The string was created based upon the following strategy:

1. **Keyword:** "value-based software engineering"
   Synonyms/adaptations: "value-based" AND "software engineering"; "value based" AND "software engineering"; "Value based software engineering"; VBSE.
2. **Keyword:** "value"
   Synonyms/adaptations: "value based"; "valuation"; "value creation".
3. **Keyword:** "software engineering"
   Synonyms/adaptations: "software"; "software development".
4. **Sub-string:** ("economics based" OR "decision making" OR economics OR "software project").

The choice of these terms is due to the fact that in the early days software companies are forced to create value along many dimensions mainly on the economic, social, cognitive, etc. [2]. Software developers also need to know the economic implications of their decisions in development process, hence analyzing economic value is considered a complex task. The terms "economics based", or "economics", as well as "decision making" in "software project" are commonly appeared in the VBSE literature that we were aware of. We have conducted an automatic search on electronic databases, which was later complemented with snowballing [18]. In relation to the electronic search, we selected articles published up until September 2020.

3.2. Databases

We included online databases that indexed each of the VBSE papers already known prior to conducting the study. In addition, there were also previous systematic literature reviews

and mapping studies that provide recommendations on the most adequate online databases to use (e.g., [19]). Based upon both, we decided to use the following databases: IEEEXplore, ACM digital library, Scopus, ScienceDirect, ISI Web of Science, and SpringerLink. These databases were selected because they have been considered as relevant ones by Dyba et al. [20] and Kitchenham and Brereton [21]. Note that although there are a few other potential databases such as EI compendex, Wiley Interscience (Wiley Online), Inspec and Kluwer as identified in [22], these databases were excluded in our mapping study due to the high degree of overlap among the databases, as reported by [19].

3.3. Study selection

Based on the guidelines presented in [15] and [23], we used the selection criteria as shown in Table 2. The main inclusion criteria were to consider any VBSE related studies, and for VBSE to be mentioned either in the title, abstract or keywords. This means that only studies that considered value aspects as per the value-based principles defined by [1] were considered. However, despite the use of a validated search string, and more strict inclusion criterion (IC 02), the study selection phase was not straightforward because in many cases we were unable to decide on whether to include or exclude a paper based solely on the paper's title, abstract, and keywords. In most cases their full text had also to be referred to, so to be sure that the paper presented a VBSE research.

Table 2. Inclusion and exclusion criteria

Inclusion criteria
IC 01 – Studies that are related to VBSE IC 02 – The title and/or abstract and/or keywords do(es) explicitly mention(s) VBSE IC 03 – Fully refereed journal and conference papers, and book chapters IC 04 – Articles/chapters written in English
Exclusion criteria
EC 01 – The publication lies outside the SE domain EC 02 – The publication is a grey literature (e.g., thesis) EC 03 – Papers not written in English EC 04 – The full text of the paper is not available EC 05 – The publication is within the SE domain but not related to VBSE

3.4. Classification scheme

To create a map of VBSE publications, we applied the general classification approach suggested by [23]. General classification refers to classifications that are used by majority of mapping studies [23]. In this mapping study, we referred to the following classifications: i) research topics, ii) publication venue, iii) research method, iv) research type, and v) contribution facet.

3.5. Data extraction

The items used for the data extraction are shown in Table 3, where we can also see which extracted data was used to help answer the RQs. The data extracted from each paper are

stored in a Google spreadsheet using the items listed in Table 3. The strategies used in extracting the data are described below:

- **Value definition:** The term "value" is searched throughout the paper to identify if there is any specific definition given, and whether the authors refer to Boehm's seminar paper [1] or VBSE book edited by [2] in determining the context of value or value-based used in the study.
- **Quality of study:** The quality of studies is rated quantitatively based on four aspects: reporting, rigor, credibility, and relevance, based on the classification of research quality proposed by [7].
- **VBSE principles and practices:** Classification of VBSE principles and practices is determined based on the agenda in [1] (e.g., VB requirements engineering, VB planning and control, etc.). We searched for specific agenda reported in the paper, however if it is not explicitly mentioned in the title, abstract or introduction Section, we inferred based on the objective(s), aim(s) and the outcomes of the study.
- **Research topic:** Articles were classified according to the SWEBOK's Knowledge Areas [24], identified based upon their keywords, main topic and focus. Major keywords (e.g., keywords and terms appeared in the title) and the dominant focus of each study are captured to identify the topics investigated. This is performed through a qualitative analysis of each primary study. The identified keywords and topics were then classified or grouped using a broader category and then we mapped these categories to the Knowledge Areas defined in SWEBOK [24] (e.g., [S99] focuses on requirements negotiation, hence classified under the Requirements Knowledge Area). The same method has been applied in [25] using the earlier version of SWEBOK.
- **Research method:** Research method is classified based on the methodologies suggested in [26, 27]. These include: Controlled Experiment, case study, survey research, ethnography, action research, simulation, prototyping, mathematical analysis, mathematical proof properties, literature review, and mixed method.
- **Research type:** Research type is identified based on the category defined in [27] and using the decision table for research type classification suggested in [23].

Table 3. Items for data extraction

Data item	RQ
Study ID	–
Value definition	RQ1
Quality of study (reporting, rigor, credibility, and relevance)	RQ2
VBSE principles and practices (according to VBSE agenda [1]), e.g., VBRE, architecting, designing and development, etc.	RQ3
Research topic (e.g., software requirements, software design, etc.)	RQ4
Research method (Controlled Experiment, case study, survey research, ethnography, action research, simulation, prototyping, mathematical analysis, mathematical proof properties, literature review, mixed methods)	RQ5
Research type (e.g., evaluation/validation/solution proposal/philosophical paper/experience report/opinion paper)	RQ6
Contribution facets (i.e., type of intervention, e.g., process, method, model, tool, or metric)	RQ7
Publication venue (e.g., conference, journal, etc.)	RQ8
Bibliographic information (title, abstract, publication year, country)	Demographics
Study context (i.e., context being studied, e.g., academic, industrial, government, organization context, etc.)	Demographics

- **Contribution facets:** Contribution facets are identified based on the type of contributions as suggested in [23], which classify contributions as a process, method, model, tool or metric. We used as basis the term mentioned in the paper when extracting the contribution facets.
- **Publication venue:** We considered peer-reviewed venues, which include journals, conferences, and workshops as per [23].
- **Study Context:** We employed the study context as per [23]. This includes academic, industrial, government, and organization context.
- **Trends of research:** This is measured by counting the number of publications per year for each VBSE agenda item proposed by [1].

4. Results

4.1. Overview

Our mapping study's search and study selection process comprised three (3) stages, as shown in Figure 2. During Stage 1, we conducted electronic searches on six (6) online databases and retrieved a total of 6536 studies. The results of our automated search process are summarized in Table 4. Results showed most of the included papers were from Scopus (54), followed by IEEEXplore (28), and ISI Web of Science (27).

In Stage 2, we selected relevant studies based on the inclusion and exclusion criteria. Out of 6536 studies, we selected 126 studies that fulfilled the criteria. In Stage 3, we conducted a backward snowballing, using the included studies as seed set, in order to manually check paper references for possible inclusion of other relevant studies. Out of 3273 references, we selected 17 studies that fulfilled our selection criteria. Hence, the final count of selected studies was 143 studies, 126 from the screened automated searches, and 17 from

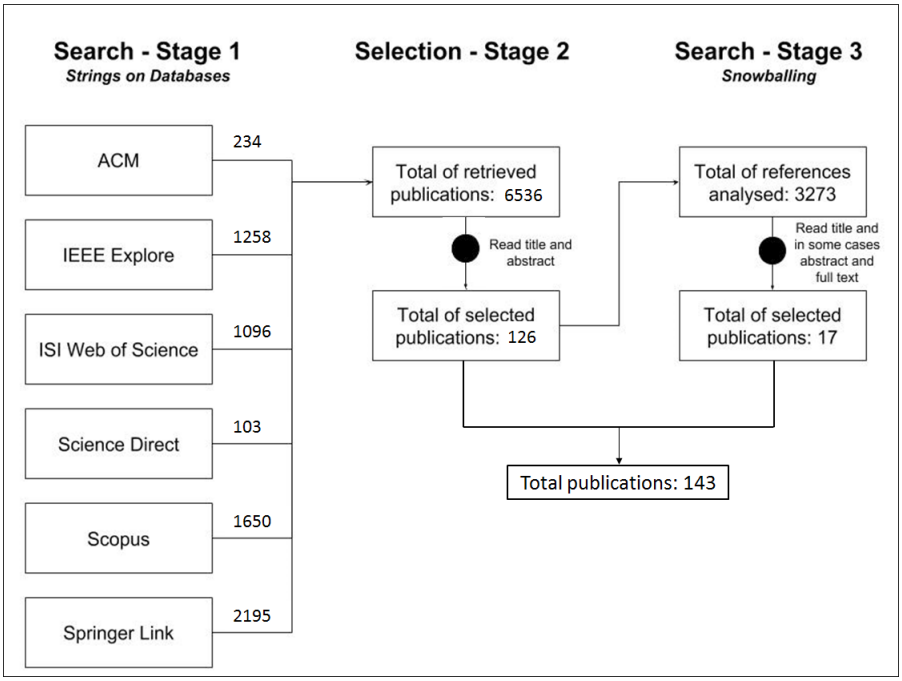

Figure 2. Study selection process

reference snowballing. These 143 studies are listed in Appendix A. Each study is identified as Sn, where n represents the study's number. Out of these 143 studies, 37 (26%) were published in journals, 88 (61%) in conference proceedings and the remaining 18 (13%) as book chapters. Figure 3 shows the number of studies that were published each year, since 2003. During the first two years (2003 and 2004) there were only five and two studies published, respectively; publication numbers peaked in 2006 with 24 studies (12 of these were book chapters in the VBSE book [2]). Overall, we have seven years (2005, 2006, 2007, 2008, 2009, 2010 and 2013) with at least 10 publications per year, followed by another six years (2003, 2011, 2012, 2014, 2015, 2017) with at least 5 publications per year. Since 2017 numbers have declined, with only two publications in 2019 and one publication in 2020.

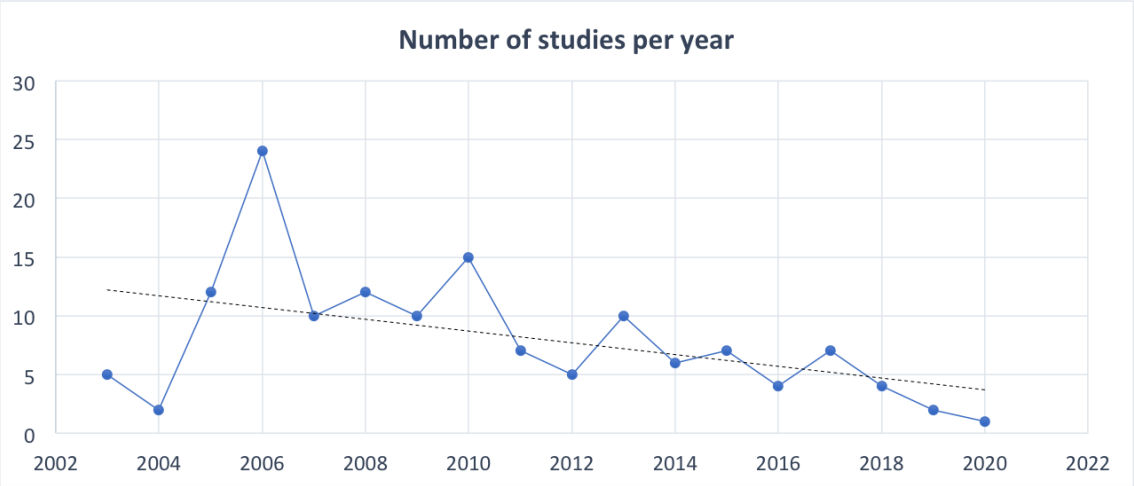

Figure 3. Line plot for the number of publications per year

4.2. Value definitions (Research question 1)

How value has been defined in the existing VBSE studies?

One important element in studying VBSE literature is to explore and to understand the notion of "value" in the primary studies. In particular, it would be interesting to know how did the studies define value, or the aspects of value the researchers are concerned with, and whether or not the notion of value goes beyond the conventional concepts of value, i.e., in terms of monetary or financial aspects. Analysis of value concepts can provide insights to

Table 4. Search results

Database	Retrieved	After removing duplicates	Excluded	Included
ACM	582	234	223	11
IEEE Xplore	1936	1258	1230	28
ISI Web of Science	1590	1096	1069	27
Science Direct	800	103	101	2
Scopus	2825	1650	1596	54
SpringerLink	2910	2195	2191	4
TOTAL	10643	6536	6410	126

the practitioners and researchers on the definition of value, including the measures used for assessing value.

In extracting the value definition used in the primary studies, we searched the term "value" throughout the paper and also identify the reference(s) related to VBSE cited in the paper, e.g., Boehm's seminal paper [1], VBSE book edited by [2]. The "value" term was first searched in the abstract, Introduction, and Conclusion, to determine if the "value" or "value-based" concepts are defined explicitly in the paper. The checking continued to the remaining Sections when no definite definition on value or value-based is found in the above mentioned Sections.

While there is no standard or commonly accepted definition of the term "value" used in the primary studies, we found that majority of the studies (85%, 122 out of 143) refer to either VBSE as defined in [1] or generally cite the VBSE book [2] when describing the context of value-based in their studies. Nonetheless, there are 15 studies (10%) that provide a clear definition of the term "value" applied in their studies (see Appendix B). In the context of VBSE studies included in this mapping study, most studies treat value as beyond the monetary or financial value. For example, value is defined as "relative worth, utility, or importance" [S9], [S20], [S49], [S143], "customer loyalty, innovation technology, cost reduction" [S25], "degree of desirability" [S77], "benefits derived from the product" [S101], "requirements fulfillment" [S110].

Although many researchers focused on the multi-dimensional perspectives of value, the economic perspective of value is undeniably important. This is because economics are considered most important in making business decisions, and as such form the basis of valuation of software assets and projects [S47]. According to Erdogmus et al. [S47], the process of determining the economic value of a product or service is not straightforward due to uncertainty in software development project. Several techniques such as decision-tree and real-options theory can be used to demonstrate how valuation can help with dynamic decision making under uncertainty. Earlier proponents such as in [8] have also promoted the concepts of economics or business value in support to decision making in software engineering.

One notable definition of value that is not related to economics, utility or functional value is given by Thew and Sutcliffe [S139]. Values in their study are defined as "personal attitudes or long-term beliefs, which may influence stakeholder functional and non-functional requirements". Values are also interpreted as "a set of issues which are frequently referred to as problematic in the RE process, such as politics, culture, sensitivities about the consequences of automation and conflicts between stakeholders" (p. 443). They mentioned that 'socio-political' issues such as emotions, values and people's feelings are often cited as problems in Requirements Engineering, hence proposed a method for analyzing such issues. They proposed a VBRE method that guides the elicitation of stakeholders' values, motivations, and emotions. Similar to [S139], we found another study [S15] that also considered value as personal beliefs and attitudes. In [S15], they referred to Schwartz's Value Theory [28], which emphasizes the profound nature of value in combination with an analysis of human motivation. The basic values are classified as: Power (authority/wealth), Universalism (equality/justice), Achievement (success/ambition), Benevolence (helpfulness), Hedonism (pleasure), Tradition (humility/ devotion), Stimulation (exciting life), Conformity (obedience), Self-determination (creativity/freedom), and Security (social order). In addition to these basic values, [S15] also referred to Holbrooks Typology of Consumer Value that defines consumer value as ''an interactive, relativistic preference experience''. Based on the

concepts of basic values and consumer values, the authors in [S15] proposed a meta-model to capture consumer preferences to be accommodated in IT system development.

Although the majority of studies applied the value-based concept as described in [1], or referred to the VBSE book [2] in general, we found 21 studies that have neither cited [1] nor [2] with regard to the "value" concept used in their studies (see Table 5). For example, in [S12], value is referred as business goals (cost, time to market, etc.) and the Return-on-Investment (ROI) technique is used to measure the business value. A total of 48 studies (see Table 5) refer to the stakeholder's value or perceived value from multiple stakeholders. Studies that focused on stakeholders' value propositions are mainly concerned about prioritizing requirements based on each requirement's perceived value, e.g., [S8], [S13], [S32]. Finally, there are 14 studies that focus on customer value or value creation in Agile (e.g., [S10], [S33]).

Table 5. List of studies – value perspectives

Value perspectives	Studies	# Studies
Studies that provide explicit definition of value	[S9], [S15], [S20], [S25], [S26], [S38], [S47], [S49], [S77], [S95], [S101], [S110], [S126], [S139], [S143]	15 studies
Studies that used value-based concept based on either Boehm's definition of VBSE [1] or VBSE book [2]	Majority of the studies (85%)	122 studies
Studies that neither cite [1] nor [2]	[S12], [S15], [S27], [S62], [S82], [S113], [S114], [S119], [S123], [S124], [S125], [S127], [S128], [S132], [S133], [S134], [S135], [S136], [S138], [S139], [S142]	21 studies
Studies that refers to or focus on stakeholder's value or perceived value	[S8], [S13], [S16], [S20], [S21], [S24], [S28], [S29], [S32], [S34], [S37], [S38], [S41], [S43], [S44], [S45], [S48], [S51], [S52], [S53], [S55], [S57], [S60], [S63], [S65], [S67], [S69], [S70], [S72], [S75], [S81], [S86], [S91], [S97], [S99], [S102], [S103], [S104], [S105], [S107], [S108], [S111], [S112], [S116], [S117], [S125], [S135], [S140]	48 studies
Studies that focus on customer/consumer value or creation of business value in Agile	[S10], [S33], [S56], [S61], [S62], [S64], [S82], [S84], [S88], [S89], [S115], [S119], [S127], [S136]	14 studies

Summary of key findings:
1. The term "value" or the concept of "value-based" has not been clearly or explicitly defined in many VBSE studies but most studies have cited either the seminal paper by Boehm [1] or the VBSE edited book [2].
2. Most of the studies defined "value" from the perspective of relative worth, or utility, as compared with economic value.
3. The measures used to evaluate or represent value is not explicitly mentioned neither described in many VBSE studies.

4. The varying notions of value concepts could potentially hinder the development of software systems, hence require collaboration with practitioners in order to implement specified values in software development.

4.3. Quality assessment (Research question 2)

What do we know about the quality of VBSE studies, particularly on the quality of reporting, rigor, credibility and relevance?

To address this research question we have used a classification of research quality proposed by [7], where the quality of primary studies is assessed based upon 11 criteria, arranged into four main aspects. This is a detailed quality criteria that does not have the high level of ambiguity of other existing proposals in Software Engineering [29]. This classification's four main aspects and the corresponding criteria are as follows:

1. Reporting – Contains three criteria that assess the quality of reporting of a study's rationale, aims and context. The three criteria are:
   a) Is the paper based on research (or is it merely a "lessons learned" report based on expert opinion)?
   b) Is there a clear statement of the aims of the research?
   c) Is there an adequate description of the context in which the research was carried out?
2. Rigor – Contains five criteria that assess the thoroughness of the "research methods employed to establish the validity of data collection tool and the analysis methods". It characterises the "trustworthiness of the findings". The five criteria are:
   a) Was the research design appropriate to address the aims of the research?
   b) Was the recruitment strategy appropriate to the aims of the research?
   c) Was there a control group with which to compare treatments?
   d) Was the data collected in a way that addressed the research issue?
   e) Was the data analysis sufficiently rigorous?
3. Credibility – Contains two criteria that assess the trustworthiness of the study's methods so to ensure that "the findings were valid and meaningful". The two criteria are:
   a) Has the relationship between researcher and participants been adequately considered?
   b) Is there a clear statement of findings?
4. Relevance – Contains one criterion that assesses the importance of "the study for the software industry at large and for the research sommunity". The criteria is: Is the study of value for research or practice?

Each of the 11 criteria was measured as "yes" or "no", and later we counted the number of "yes" for each of the four main aspects, for each of the 143 studies. This means that the range of values for each main aspect was as follows: Reporting (0–3); Rigor (0–5); Credibility (0–2); and Relevance (0–1). We have also associated labels with these values, which are used below while discussing the results, and also in the Discussion Section, when comparing quality to other aspects also investigated herein. The labels used are as follows:

- Reporting: 0 – Unsatisfactory; 1 – Acceptable; 2 – Good; 3 – Very Good.
- Rigour: 0 – Unsatisfactory; 1 – Poor; 2 – Acceptable; 3 – Good; 4 – Very Good; 5 – Excellent
- Credibility: 0 – Unsatisfactory; 1 – Acceptable; 2 – Good; 3 – Very Good.
- Relevance: 0 – Unsatisfactory; 1 – Very Good.

The 11 criteria for each of the 143 papers are measured, and the summary results are shown in Figure 4. It shows that the majority of studies presented very good reporting quality and relevance. The results for studies' rigour were somewhat mixed, with the largest number of studies showing poor rigour, followed by very good and unsatisfactory rigour. As for credibility, the majority of studies presented acceptable credibility, followed by good credibility. Two quality aspects – relevance and reporting, provided results showing that a clear majority of studies were judged to be very good. However, rigor and credibility do not present the majority of studies with higher quality. In fact, most studies were judged to present acceptable credibility only, and poor credibility.

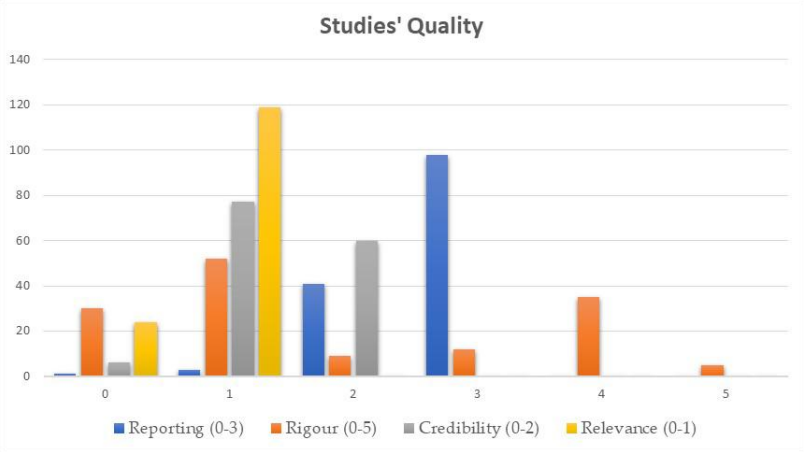

Figure 4. Studies' quality as per four main quality aspects

The 11 quality criteria used to assess the studies' quality is shown in Figure 5. It shows that out of the five criteria used to assess rigor, only criterion 7 presents a "Yes" for most studies. The worst result was for criterion 6, which represents an important aspect to manage when carrying out formal experiments. Furthermore, criterion 8 – data analysis used, also seemed to lack rigor for a large number of studies (102), and criterion 5 also showed that many studies (91) lacked the use of an appropriate recruitment strategy. Such poor results, particularly for criteria 5, 6 and 8, were also observed by [7], when assessing 33 empirical studies of agile software development. Their best results were also obtained for Reporting and Relevance. However, unlike the studies investigated by [7], the ones included in this mapping study were characterised by a large number of studies that did not carry out empirical investigations. Instead, they provided detailed examples – proof of concept, about the solutions they were proposing (e.g., prioritization technique, tool). A total of 44 studies (31%) were proof of concept studies. Another set of 20 studies (14%) presented proposals without a detailed example and no empirical evaluation or even proof of concept is provided.

We also wanted to assess whether there were statistically significant associations between the four different quality aspects; therefore we carried out a Pearson's test, with Bonferroni correction, to measure the strength of association between the four aspects. The analysis was carried out using Stata, with $\alpha = 0.05$. Results are displayed in Figure 6, in which there are three values shown per pairwise correlation: The first is the correlation coefficient (an asterisk showed that the coefficient is statistically significant); the second is the p-value of the test, and the third is the sample size used. The results indicate that there are statistically significant positive associations between the four quality aspects; however the

Figure 5. Detailed 11 quality criteria, arranged according to four aspects

	repor~03	rigour05	credi~02	relev~01
reporting03	1.0000			
	143			
rigour05	0.4125*	1.0000		
	0.0000			
	143	143		
credibili~02	0.3524*	0.7313*	1.0000	
	0.0001	0.0000		
	143	143	143	
relevance01	0.3893*	0.4259*	0.3003*	1.0000
	0.0000	0.0000	0.0016	
	143	143	143	143

Figure 6. Detailed 11 quality criteria, arranged according to four aspects

highest correlation coefficient relates to the relationship between Rigor and Credibility (0.7313). Therefore, the higher the credibility of a paper, the higher its rigor, and vice-versa. The second highest correlation coefficient (but much lower than the highest) was obtained for Relevance and Rigor (0.4259); thus the higher the Relevance, the higher the Rigor, and vice-versa. It is important to note that the highest correlation coefficient was not given for Rigor and Relevance, which, in our view, suggests that the use of more detailed

quality measures, such as the one employed herein, rather than solely Rigor and Relevance, provides a better and more detailed understanding of studies' quality.

Summary of key findings:
1. The studies' quality criteria by Dybå and Dingsøyr [7], when applied to the 143 studies in this mapping study, showed that most studies presented good quality of reporting and relevance, acceptable credibility and poor rigor.
2. Many of the studies published within the period 2003 to 2016 were either proof of concept, or advocacy research-type papers; however since 2017 all studies presented evidence obtained by means of empirical investigations.
3. The Pearson's correlation analysis test showed a statistically significant high positive association between rigor and credibility.
4. Results indicated that empirically-based studies, and with higher quality in terms of rigor and credibility, are needed.

4.4. SE principles and practices (Research question 3)

What are the SE principles and practices investigated so far in VBSE, and how has this changed over time?
To answer this research question we used as basis the classification suggested in VBSE Agenda for existing and emerging SE principles and practices [1] (requirements engineering, architecting, design and development, verification and validation, planning and control, risk-management, quality management, people management, and Theory of VBSE), plus an additional four practices not included in the original agenda (value-based decision-making, software process, value creation and a fourth category called "Other", i.e., studies looking at general aspects of VBSE). These four additional classifications were added in order to better characterize some of the selected studies, and in line with their research descriptions. Figure 7 shows the number of studies arranged per SE principles and practices, and per the three different periods being covered (2003 to 2008; 2009 to 2014; 2015 to 2020), and

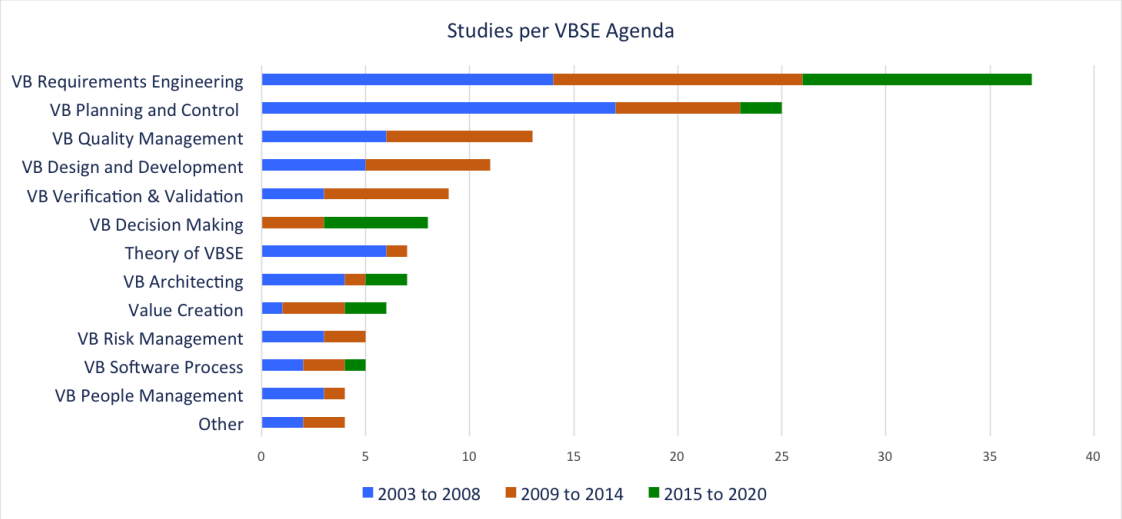

Figure 7. Studies per VBSE agenda

Table 6 provides further details relating to which studies belong to a given category, also arranged according to the same time periods. First, we will elaborate upon the overall results prior to discussing whether, and how, trends have changed over time.

Table 6. VBSE agenda

VBSE Agenda	Count	Paper ID
VB Requirements Engineering	37	2003 to 2008 (14) [S18], [S19], [S21], [S28], [S61], [S79], [S86], [S97], [S99], [S101], [S104], [S107], [S109], [S126] 2009 to 2014 (12) [S7], [S8], [S13], [S14], [S35], [S60], [S62], [S65], [S72], [S116], [S119], [S130] 2015 to 2020 (11) [S15], [S55], [S56], [S103], [S124], [S135], [S136], [S137], [S139], [S140], [S142]
VB Planning and Control	25	2003 to 2008 (17) [S20], [S23], [S24], [S30], [S32], [S39], [S40], [S70], [S75], [S92], [S51], [S68], [S81], [S96], [S105], [S111], [S134] 2009 to 2014 (6) [S36], [S43], [S95], [S118], [S123], [S131] 2015 to 2020 (2) [S108], [S141]
VB Quality Management	16	2003 to 2008 (9) [S16], [S17], [S22], [S31], [S41], [S53], [S57], [S74], [S94] 2009 to 2014 (7) [S10], [S11], [S113], [S114], [S125], [S27], [S91] 2015 to 2020 (0) none
VB Design and Development	11	2003 to 2008 (5) [S52], [S64], [S66], [S106], [S129] 2009 to 2014 (6) [S25], [S37], [S63], [S77], [S78], [S132] 2015 to 2020 (0) none
VB Verification and Validation	09	2003 to 2008 (3) [S46], [S71], [S100] 2009 to 2014 (6) [S6], [S59], [S73], [S83], [S93], [S98] 2015 to 2020 (0) none
VB Decision Making	08	2003 to 2008 (0) none 2009 to 2014 (3) [S84], [S88], [S90] 2015 to 2020 (5) [S34], [S54], [S112], [S117], [S143]
VB Architecting	07	2003 to 2008 (4) [S12], [S58], [S80], [S85] 2009 to 2014 (1) [S89] 2015 to 2020 (2) [S122], [S133]
Theory of VBSE	07	2003 to 2008 (6) [S1], [S48], [S49], [S50], [S69], [S76] 2009 to 2014 (1) [S4] 2015 to 2020 (0) none
Value Creation	06	2003 to 2008 (1) [S127] 2009 to 2014 (3) [S9], [S110], [S115] 2015 to 2020 (2) [S33], [S138]
VB Software Process	05	2003 to 2008 (2) [S42], [S44] 2009 to 2014 (2) [S5], [S87] 2015 to 2020 (1) [S120]
VB Risk Management	04	2003 to 2008 (3) [S29], [S47], [S102] 2009 to 2014 (1) [S121] 2015 to 2020 (0) none
VB People Management	04	2003 to 2008 (3) [S2], [S45], [S128] 2009 to 2014 (1) [S38] 2015 to 2020 (0) none
Other	04	2003 to 2008 (2) [S26], [S67] 2009 to 2014 (2) [S3], [S82] 2015 to 2020 (0) none

Results show that **VB Requirements Engineering (RE)** has been the mostly investigated principle and practice in VBSE, contributing with 37 studies (25.3%). Majority of the studies (17 studies) proposed a value-based method, or approaches for requirements prioritization. Other value-based approaches proposed include the areas of requirements elicitation, requirements tracing, RE process, requirements negotiation, and tool support selection. The second most investigated principle and practice in VBSE research is **VB Planning and Control**, with 25 studies (17.1%). Most studies (9 studies) under this category proposed value-based approach to support software project planning. The rest of the studies proposed value-based methods for software release planning, managing value delivered to stakeholders, value-based technique to better prioritize stakeholders' value, value-based approach to measure productivity, planning for software traceability, value assessment for software reuse, planning for measurement to support decision-making process, and value-based approach to determine an optimum software assurance investment.

The additional three (3) principles and practices that also received significant attention in VBSE research are VB Quality Management (16 studies), VB Design and Development (11 studies) and VB Verification and Validation (9 studies). Research in **VB Quality Management** mainly focused on software processes' quality aspects (4 studies). The remaining studies investigated the levels of alignment between key stakeholders on software quality aspects, value aspects of software quality assurance, tailoring the value-based software quality achievement process to different business cases, software quality investment, and assessment of quality processes. Research related to **VB Design and Development** involves techniques and approaches to ensure value-considerations are integrated into the software's design and development [1]. Three (3) out of 11 studies proposed value-based approach to support software component markets. The remaining studies proposed a design technique used to estimate the value of a design strategy, an approach to develop decision support systems, incorporating customers' value in the process of partitioning hardware and software for embedded system, managing inconsistencies in software development, and value-based technique to evaluate software designs. **VB Verification and Validation (V&V)** has been researched in nine (9) studies. Two (2) of these focused upon prioritization strategies to improve software testing cost-effectiveness. Others have proposed a value-based software testing method to better align investments with project objectives and business value, enhancement of V&V process, coverage measurement tool in software system testing, and software evolutionary testing framework using genetic algorithms. Two (2) experimental studies on VB V&V compared the performance of value-based review (VBR) with the traditional value-neutral checklist based reading approach.

We identified eight (8) studies on **VB Decision-Making** as an emerging research area in VBSE. Two (2) studies explored feature usage measures to support the decision-making process. Two studies introduced a VALUE framework to estimate the value associated with stakeholders' decisions. They also developed a Value tool to support the decision-making process. The other studies proposed a software value map for making decisions about product management and development, and empirical studies to validate models for estimating value of decisions, and assessment of a Web-based tool for value-based decision-making.

There are seven (7) studies found related to **VB Architecting**. Three (3) of these studies focused on value-based approach for documenting design-decisions rationale to support software architecture design. The remaining studies introduced: a customer-centric value for assessing system architecture investment, a lightweight value-based architecture evaluation, a value discovery method in the context of Big Data design, and a method to evaluate diversification of software architecture for software sustainability.

Three of the seven (7) studies classified under **Theory of VBSE** described the $4 + 1$ theory. One study [S4] made a proposal for extending the VBSE theory. The remaining three (3) studies present the VBSE agenda and the seven VBSE elements. **Value Creation** category comprises six (6) studies, in which the majority (4 studies) focuses on customer value creation in Agile context, while the others proposed a new definition of value, and an empirical study on how user perceived value impacts user loyalty for software product. Five (5) studies under the **VB Software process** mostly investigated value factors that can impact software development process and the factors were later used in building a framework for software process tailoring. Others had introduced vaue-based software process model for Europe, and a value-based set of processes for Components-Off-The-Shelf (COTS)-based applications.

Our results showed that VB Risk Management and VB People Management are the least investigated VBSE principles and practices (four (4) studies respectively). Studies related to **VB Risk Management** have proposed: valuation of software initiatives under uncertainty to help with decisions at the project level, a value-based process to manage requirements-related risks, a model to identify risk in architectural mismatches in component-based system development, and a method to assess uncertainties in software project. The four (4) studies classified under **VB People Management** described four different aspects: value-based knowledge management to support learning in software companies, value-based approach for managing architectural knowledge, collaborative process to facilitate stakeholders' involvement, and stakeholder value as a means to understand conflicts in software development.

The remaining four (4) papers in the **"Other"** category are general VBSE papers covering a framework to identify value of new innovation idea, applicability of Lean Six-sigma principles to be embedded in VBSE process, applications of machine learning methods in VBSE, and pedagogical game for teaching VBSE to students. Overall findings showed that most VBSE studies had focused on the early phases of software engineering activities, i.e., requirements engineering, and planning and control. While various value-based approaches and solutions have been proposed (as described above), initiatives to perform measurement of value in VBSE studies require further addressing. This is because such measurement is needed in various SE activities as a follow up on the generation of value.

When we look at the trends over the three different periods (in Figure 7), we see that the only VBSE principles and practices remained with a similar number of publications over all three periods has been VB Requirements Engineering. VB Planning and Control had the largest number of publications over the period 2003 to 2008, but then dropped by less than half over the next period (2009–2014) and down to two papers between the period 2015–2020. VBSE principles and practices not investigated during the most recent period (2015–2020) are VB Quality Management, Theory of VBSE, VB Design and Development, VB Verification and Validation, VB Risk Management, VB People Management, and Other. Such lack of recent studies in areas that are still relevant within SE suggests possible research gaps that could be investigated by the VBSE community. VB Decision Making only emerged, as far as publications are concerned, over the two most recent periods, with an increase in publications over the most recent period (2015–2020). Both Theory of VBSE and VB People Management had by far their largest contribution in number of publications during the first period 2003–2008.

The overall trends of publications based on VBSE principles and practices can be seen in a bubble plot (see Figure D1 in Appendix D). In the bubble plot, the size of a bubble indicates the amount of papers published and the number near a bubble represents the

number of publications. Based on the number of publications, the trend indicates that there is a constant interest in VB Requirements Engineering research, followed by VB planning and control, and Value creation. VB quality management is also an area that had some interest up to 2014. Other principles and practices that had publications for at least four years (not necessarily consecutive years), but later discontinued, are VB verification and Validation, VB Software Process, VB design and development, VB architecting, Theory of VBSE, and Other. As previously mentioned, the only emerging area is VB decision making, with publications since 2013. Finally, principles and practices that had publications for three years (not necessarily consecutive years) are VB risk management and VB people management.

We observed that the last year in which there were papers co-authored by Barry Boehm was 2013, and from that point onwards there was a decline in the number of papers that considered value aspects as per the value-based principles defined by [1]. This was observed very clearly for those seven principles and practices abovementioned, for the most recent period (2015–2020). However, this does not necessarily mean that VBSE is not important, or that value-based research in SE ceased to receive attention. Some of the VBSE principles and practices may have been adopted by the Lean and Agile Software development communities for example through a continuous value delivery practice as highlighted in [30], while others may perhaps be the focus of some research that does not explicitly reference Barry Boehm's work in VBSE. An additional point to stress here is that despite our use of number of publications to suggest research gaps, the number of publications is not the only factor to identify gaps. There is a need to determine (a) is research needed where there are few publications and (b) are areas with many publications still supported by credible evidence, considering studies' quality. With regard to the latter point, it was discussed in detail by RQ2.

Summary of key findings:
1. The main emphasis of VBSE research is placed on the early phases of software engineering, i.e requirements engineering and software planning/control.
2. Within RE, the main concern of investigations was placed on using value in the prioritization of requirements.
3. Not much is known on how values can be incorporated in SE practices to analyze, prioritize and mitigate risks that occur in software project (VB Risk Management).
4. There are lack of studies in the areas of VB Risk Management and VB People Management.

4.5. Research topics (Research question 4)

What are the most investigated research topics in VBSE, and how has this changed over time?

To answer this research question, we referred to the twelve (12) Knowledge Areas identified in SWEBOK [24] (e.g., software requirements, software design, software construction, software testing, etc.). The result showed that Software Requirements and SE Management were the two topics that have been actively researched in VBSE (34 and 30 studies respectively). Within the Software Requirements category, most studies (17 out of 34 – 50%) focused on issues related to requirements prioritization. Studies within the SE Management category focused on different management aspects mainly decision value/analysis (6 studies), and

product value estimation and planning and control (4 studies each). Our findings showed that very few studies fell under the SE Professional Practice (3 studies), Software Maintenance (2 studies), and Software Construction (1 study) Knowledge Areas. It is interesting to note that there are no studies available particularly addressing the SWEBOK Knowledge Areas of Configuration Management, Computing and Mathematical Foundations.

With regard to changes in VBSE research topics over time (see Figure 8), six topics showed some consistency in the number of studies for the first two time periods (2003 to 2008; 2009 to 2014): Software Requirements, Software Quality, SE Process, SE Models and Methods, SE Management and Engineering Foundation. However, except for Software Quality and Engineering Foundation, which did not have any studies published within the period 2015–2020, all four remaining topics dropped their number of studies published within the period 2015–2020 by half or even less than half. The SE Professional Practice topic only had studies published within the period 2003–2008, and Software Construction only had one study published, and in the period 2015–2020. The topic Software Maintenance only had studies published within the two most recent periods. Despite the drop in the number of studies over the most recent period 2005–2020, there are nine topics, out of the 12 topics, which had at least one study published over the most recent period. Overall, the two topics Software Requirements and SE Management showed the highest numbers of publications for all the three periods covered, thus suggesting a continued interest from the research community in these two topics. The detailed list of research topics addressed by our primary studies is available in Appendix C.

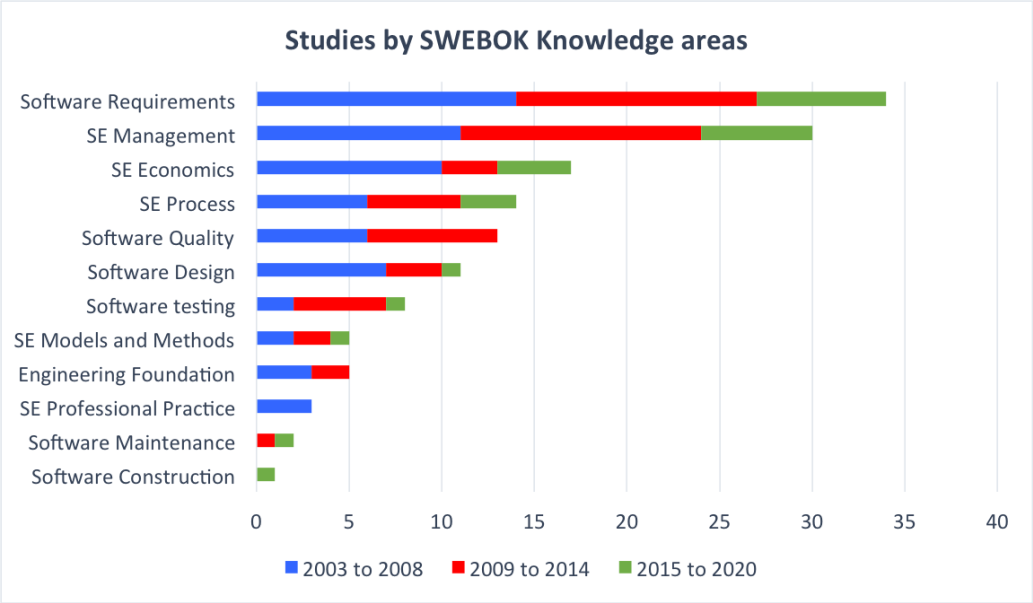

Figure 8. Research topics trend over time

Summary of key findings:
1. Most of the VBSE research topics fall within the area software requirements, mainly requirements prioritization.

2. Majority of the topics (9 out of 12) had at minimum one publication over the most recent period, suggesting that the topics are still active (despite low number).
3. Topics related to software maintenance, software construction and SE Professional practice have received less attention despite being important areas in SE.

4.6. Research methods (Research question 5)

What are the research methods used in VBSE studies and how many studies looked at each method (e.g., case study, experiments, survey, etc.)?
This research question aims to identify the research method(s) employed in the primary studies included in this mapping study. We used the classification of research methods as reported in [26, 27]. The bar chart in Figure 9 shows the distribution of studies by the research method.

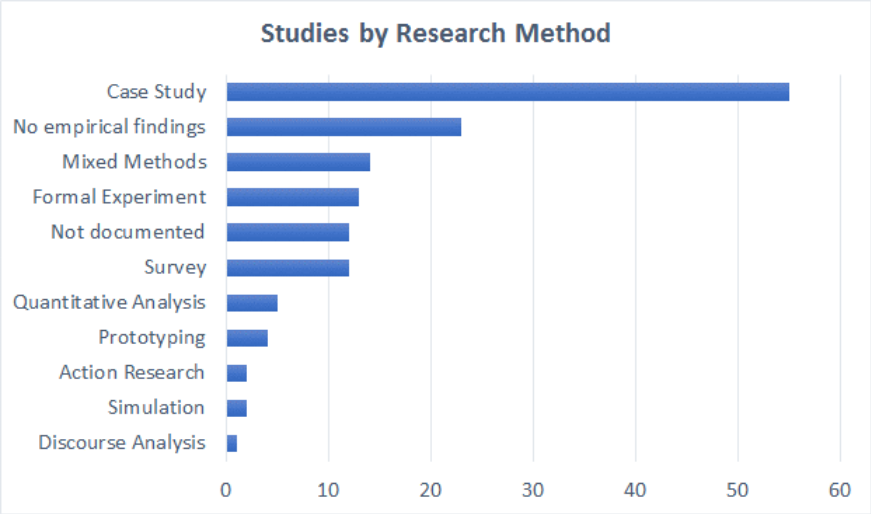

Figure 9. Studies by research method

Our analysis showed that most studies (38%, 55 out of 143) were conducted using a case-study methodology. Forty-two (42) of these studies reported their case study within an organizational context (e.g., defence agency, software organization, startup company and large company such as Ericsson). After case studies, the second largest category (23 studies) related to studies that did not report empirical findings. They proposed solutions without empirical validation or evaluation (e.g., [S6], [S46], [S62]). Next, we had, Mixed methods (14 studies), Formal experiment (13 studies), Not documented (12 studies), and Survey (also 12 studies). Finally, the last five categories with low number of studies were Quantitative analysis, Prototyping, Action research, Simulation, and Discourse analysis.

The breakdown of studies according to the study context is available in Table 7. Studies that have used more than one research method were classified under the mixed-methods category. Case study is by far the research method used the most, and most of the case studies were carried out within an organizational setting. Similar to case study research, mixed methods and survey studies were also commonly conducted in organizational context. Studies that have used formal experiments mostly conducted their research in academic context. A total of 23 studies did not report any empirical findings but have documented

their context; most were conducted within an organizational context. Finally, there are a small number of studies that performed simulations, discourse analysis, and action research. Analysis based on VBSE Agenda indicated that, except for VB quality management, VB people management, and VB decision making, all the other principles and practices had at least one study that had no empirical findings. This corresponds to 16% of the included studies, which, in our view suggest that future research should focus upon widening the number of studies with empirical investigations within the context of VBSE.

Summary of key findings:
1. Most of the VBSE research were conducted using case-study methodology.
2. Research conducted in industrial and organizational context commonly applied case study, mixed-method, and survey methodology.
3. Formal experiments are mostly conducted in academic setting.
4. More empirical studies are needed to validate the proposed solutions; particularly there is a lack of experimental study in industrial or organizational setting.

4.7. Research types (Research question 6)

What are the research types that these studies apply (e.g., validation/evaluation/solution proposal, etc.) and how many studies looked at each research type?

In answering this question, we referred to the existing types of research approaches as suggested by [27] for making the classification. The number of studies identified for each research type is depicted in Figure 10. Our analysis showed that most studies (42 out of 143, 29%) proposed solution technique(s) without any empirical validation or evaluation. Second came studies (39 studies) that presented solution proposal together with the validation strategy. Studies that performed evaluation and validation comprised 14% and 13%, respectively. The remaining studies, less than 5% each, were categorized into philosophical paper, solution proposal and evaluation, experience report, opinion, and literature review. The breakdown of studies for each research type can be seen in Table 8.

An overview of publications across the two dimensions VBSE agenda and research type shows that there are four research types that have been used the most: i) Solution proposal, employed in 42 studies that focused on VB design and development, VB planning and control, and VB requirements engineering, ii) Solution proposal and Validation, used by 39 studies, mainly on VB requirements engineering and VB planning and control, iii) Evaluation, used in 20 studies, mainly from the VB requirements engineering topic, and finally, iv) Validation, employed in 18 studies of VB requirements engineering as well. The two research types used the least were Literature review (1 study) and Opinion paper (3 studies).

Majority of the evaluation and validation studies (25 studies) were conducted in organizational context, followed by industrial (9 studies) and academic settings (2 studies). The research methods employed for evaluation studies consist of survey, case study, and mixed-method, whereas for validation studies, most (10 studies) were conducted using case-study method, followed by experiment (4 studies) and mixed-method (3 studies). We also found that experimental type of studies have only been used in validation studies. For example, in [S73], the experiment involving graduate software engineering team project course was conducted to compare the effectiveness of value-neutral and the proposed value-based artifact priorization process.

Table 7. Breakdown of studies by research method

Research Method	Study Context					
	Organization	Academic	Industry	Govt	Not reported	Total
Case Study	[S8], [S9], [S10], [S14], [S18], [S19], [S21], [S23], [S24], [S30], [S31], [S39], [S40], [S41], [S43], [S50], [S53], [S56], [S60], [S61], [S68], [S70], [S72], [S76], [S79], [S83], [S84], [S86], [S88], [S89], [S91], [S93], [S98], [S103], [S105], [S115], [S117], [S119], [S121], [S122], [S123], [S131]	[S29], [S95], [S109], [S34]	[S102], [S108], [S130], [S142], [S140], [S143]	–	[S17], [S64], [S133]	55
No empirical findings	[S6], [S13], [S20], [S25], [S42], [S48], [S51], [S59], [S69], [S81], [S82], [S85], [S110], [S118], [S120], [S132]	[S26]	[S49], [S101], [S106]	–	[S46], [S47], [S62]	23
Mixed Methods	[S57], [S65], [S74], [S90], [S99], [S100], [S112]	[S136], [S139], [S141]	[S54], [S104]	–	[S1], [S16]	14
Formal Experiment	[S11], [S12], [S58]	[S35], [S71], [S73], [S80], [S135], [S137], [S138]	[S116]	–	[S38], [S92]	13
Not documented	[S3], [S4], [S5], [S32], [S44], [S113], [S114], [S127]	–	[S129]	–	[S37], [S94], [S96]	12
Survey	[S7], [S27], [S28], [S125]	[S15], [S33]	[S55], [S97], [S107], [S126], [S128]	[S87]	–	12
Quantitative Analysis	[S22], [S75], [S134]	[S124]	–	–	[S66]	5
Prototyping	[S36], [S63]	[S67]	[S78]	–	–	4
Action Research	[S2]	[S45]	–	–	–	2
Simulation	[S52], [S111]	–	–	–	–	2
Discourse Analysis	–	–	[S77]	–	–	1

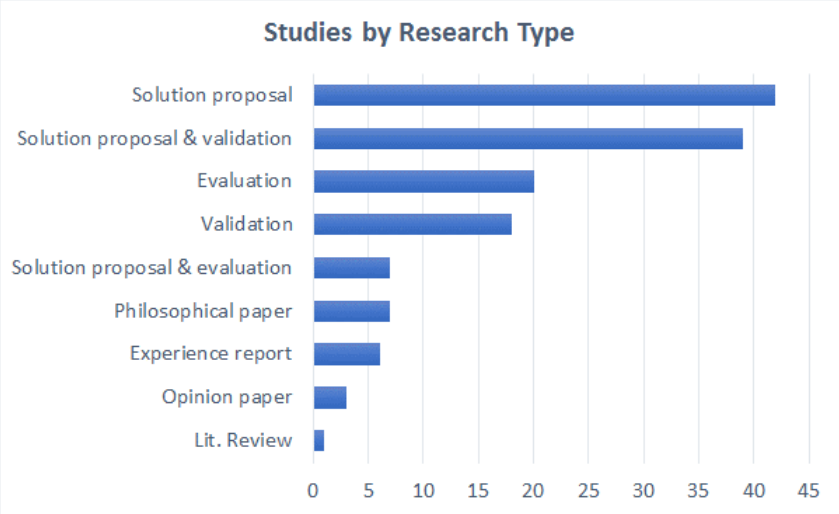

Figure 10. Studies by research type

Table 8. Breakdown of studies by research type

Research Type	Studies	Total
Solution Proposal	[S2], [S3], [S5], [S8], [S13], [S20], [S22], [S25], [S32], [S36], [S37], [S42], [S44], [S45], [S46], [S51], [S52], [S54], [S59], [S62], [S63], [S64], [S66], [S67], [S74], [S78], [S81], [S92], [S101], [S106], [S110], [S111], [S118], [S120], [S124], [S127], [S129], [S132], [S134], [S137], [S140], [S142]	42
Solution Proposal and Validation	[S6], [S11], [S12], [S14], [S15], [S17], [S18], [S23], [S24], [S29], [S31], [S35], [S38], [S39], [S40], [S41], [S43], [S68], [S71], [S76], [S79], [S80], [S83], [S84], [S86], [S93], [S95], [S96], [S102], [S103], [S109], [S122], [S123], [S133], [S135], [S136], [S139], [S141], [S143]	39
Evaluation	[S7], [S10], [S16], [S27], [S28], [S30], [S33], [S55], [S87], [S88], [S91], [S97], [S104], [S107], [S108], [S115], [S119], [S125], [S126], [S128]	20
Validation	[S34], [S53], [S56], [S57], [S58], [S60], [S61], [S65], [S70], [S72], [S73], [S75], [S105], [S112], [S116], [S117], [S121], [S138]	18
Philosophical Paper	[S1], [S47], [S48], [S49], [S50], [S69], [S77]	7
Solution Proposal and Evaluation	[S9], [S19], [S21], [S89], [S90], [S98], [S100]	7
Experience Report	[S4], [S85], [S113], [S114], [S130], [S131]	6
Opinion Paper	[S26], [S82], [S99]	3
Lit. Review	[S94]	1

Summary of key findings:
1. The most common research type in VBSE is solution proposal (comprised 61% of the studies), and almost half of these studies did not perform any empirical validation or evaluation.
2. Research methods used for empirical evaluation studies are mainly survey and case study.
3. Evaluation studies performed in industrial context have used survey as their research method.

4. Majority of the studies proposed solution without empirical validation or evaluation. This implies that there is a lack of maturity in implementing the solutions in practice and lack of evaluation involving practitioners in the real-world industrial context.

4.8. Contribution facets (Research question 7)

What contribution facets do they provide (e.g., process, method, model)? The contribution facets for each study were classified according to the contribution types suggested by [23] (see Figure 11). The facets herein referred to the contribution type or the kind of intervention being studied such as the process, method, tool, metric or model [23]. Results showed that most of the contribution facets were provided as methods (32 studies, 22.3%) followed by processes (31 studies, 21.6%), and models (24 studies, 16.7%). Next we had Frameworks (11 studies, 7.6%), Other/Tool/Metric all with 8 studies each (5.6%), Techniques (7 studies, 4.8%), Model and Tool (6 studies, 4.2%), Process and Tool/Method and Metric both with 3 studies each (2%), and finally Metric and Tool with 2 studies (1.3%).

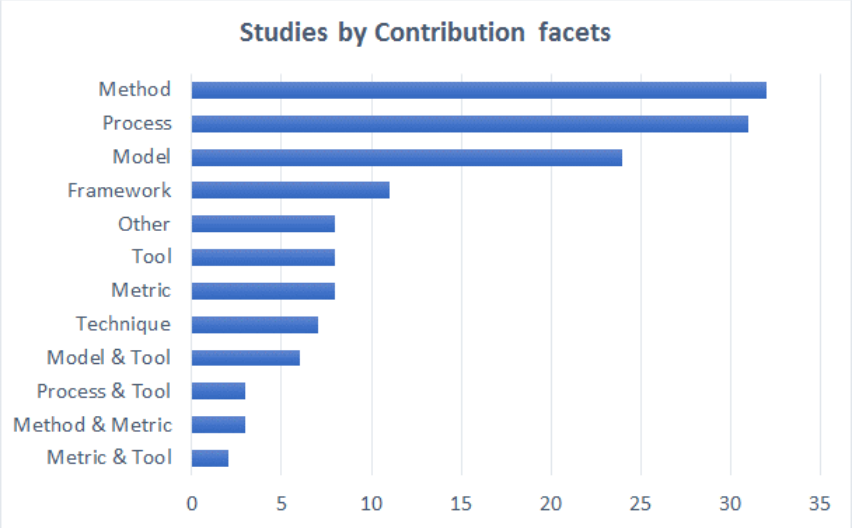

Figure 11. Contribution facets

Table 9 shows the breakdown of studies for each contribution facet. Note that while performing the mapping, we did our best to use the same facet term as identified in the paper. For example, in [S14] and [S70], the authors used the term "technique" to specify their contributions; hence, we categorized them under the "technique" category, and so on. In some studies, the authors described more than one type of contribution (e.g., [S74], [S65]). For example, [S74] proposed a quality model and a tool known as ODC-COQUALMO (Orthogonal Defect Classification COnstructive QUALity MOdel) to decompose the defect types into more granular ODC categories. Therefore, we classified this study under the model and tool category. A total of 31 studies offered a variety of solutions for improving SE processes by incorporating VBSE elements. [S11] for instance proposed an approach to transforming value-neutral processes into value-based software development processes, while [S42] presented value-based processes for COTS-based applications. A method usually

has a more specific goal and a narrow purpose or research question [31]. Most of the studies that suggested a method, focused on the effort to support requirements prioritization for elicitation and reconciliation of stakeholder value propositions. For example, [S8] put forward a method to prioritize requirements using decision theory, whereas [S24] presented a prioritization method (impact estimation) to better capture explicit stakeholder value and to cater for multiple stakeholders.

Table 9. Breakdown of studies by contribution facets

Contribution Facets	Studies	Total
Method	[S2], [S8], [S9], [S12], [S13], [S20], [S21], [S24], [S32], [S33], [S35], [S37], [S39], [S40], [S55], [S64], [S66], [S80], [S83], [S91], [S93], [S96], [S99], [S10], [S94], [S116], [S120], [S122], [S126], [S133], [S139], [S142]	32
Process	[S7], [S11], [S19], [S29], [S41], [S42], [S44], [S45], [S50], [S53], [S58], [S61], [S82], [S97], [S86], [S73], [S79], [S71], [S76], [S81], [S30], [S27], [S104], [S105], [S107], [S110], [S115], [S119], [S123], [S125], [S131]	31
Model	[S17], [S18], [S22], [S23], [S34], [S48], [S51], [S56], [S62], [S68], [S69], [S75], [S89], [S92], [S101], [S106], [S109], [S111], [S112], [S118], [S132], [S134], [S136], [S138]	24
Framework	[S3], [S5], [S15], [S43], [S46], [S59], [S77], [S85], [S31], [S16], [S141]	11
Tool	[S6], [S36], [S60], [S67], [S103], [S117], [S129], [S143]	8
Metric	[S87], [S88], [S90], [S108], [S113], [S114], [S127], [S128]	8
Technique	[S14], [S70], [S52], [S124], [S135], [S137], [S140]	7
Model and Tool	[S54], [S57], [S63], [S74], [S78], [S102]	6
Method and Metric	[S95], [S98], [S121]	3
Process and Tool	[S100], [S84], [S38]	3
Metric and Tool	[S65], [S72]	2

The contribution facet categorized as model refers to the abstract classification or model of a problem or topic, rather than to a specific tangible way of solving a problem [31]. There appears to be a number of studies presenting models – a value-driven (V2) model to elicit customers' value from requirements analysis, ROI model (iDAVE) to estimate future investment on software dependability, and value-based software assurance model to assess relative payoff of value-based vs. value neutral testing, to name a few. Our results also showed that several studies proposed a framework as their contribution facet. A framework differs from a method in the sense that it represents a detailed methodology that may include several methods, in addition to having a wider purpose and focusing on several research questions or areas [31]. An example of such a framework is the Value Elicitation Framework, proposed by Murtaza et al., 2010 [S43], which aims at facilitating the selection and application of value elicitation techniques in a project lifecycle. Zhang (2013) [S59] also describes a value-based framework that focuses on test data generation through genetic algorithms and helps prioritize decisions in the testing process.

Further, results also highlighted that some studies developed a tool for evaluating or validating proposed concepts or solutions related to processes, models, and metrics. As an example, Madachi et al., 2007 [S100] developed a software risk advisory tool using ODC COQUALMO quality model to optimize V&V processes for NASA flight projects. The "Other" contribution facets comprised a wide array of approaches that have been recommended in the identified studies, which did not relate to process, method, tool, technique, model, framework, or metric. As an example, [S4] presented an analysis of

software implementation projects for assessing the applicability of a value-based approach. When analyzing based on VBSE research agenda and research type, we found the following main aspects:

- Most studies in VB Requirements engineering used the research types Solution Proposal and Validation, Evaluation, Solution Proposal, and Validation. Regarding their types of contribution, the highest number of studies contributed towards Methods (10 studies), Processes (9 studies), Models (6 studies), and Techniques (5 studies).
- Most studies in VB planning and control employed the research types Solution Proposal and Validation and Solution proposal. Their main types of contribution were towards Models (8 studies), Methods (6 studies), and Processes (5 studies).

The results also suggest a relationship between the use of research types Solution Proposal, Solution Proposal and Validation, Evaluation and Validation with focused contributions towards Methods, Processes, Techniques and Models (see bubble plot in Appendix D). This is also supported by many other research facets and types, such as VB quality management, VB Software Process, VB design and development, VB Architecting, and Value creation. The contribution facets the least investigated were Method and Metric (3 studies), Process and Tool (3 studies), and Metric and Tool (2 studies).

Summary of key findings:
1. The contribution facet of VBSE research is mostly of method type.
2. Most research efforts is spent on value-based requirements engineering and planning and control with contributions mainly on methods, processes, and models.
3. Despite the higher number of studies contributed towards methods, processes and models, their contributions appear as proposed solution, and yet to be evaluated.
4. Contributions in terms of the proposed method, process, and model need to be supported by tools for practical use, and metrics for evaluation or measurement.

4.9. Publication venues (Research question 8)

RQ 8: What are the publication venues for VBSE research?
Most of the included studies were published in conference proceedings (88 out of 143, 61%), followed by journal articles (37 studies, 26%), and book chapters (18 studies, 13%) (see Figure 12). In relation to the book chapters, twelve (12) of these were published in the VBSE book [2]. For journal articles, we found 37 journal articles published in 22 venues.

Table 10 shows the journals where at least two VBSE studies were published. As can be seen, most journal articles were published in IEEE Software (7 out of 37 journal articles), followed by Information and Software Technology (4 articles), and the Journal of Software-Evolution and Process (3 articles). Next, there are four journals that have published 2 articles each and there are another 15 journals that have published one article each.

Table 11 lists the conferences where most primary studies were published. We only list venues in which more than one primary studies were published. In total, our review included 88 conference papers, and 42 venues where only a single primary study was published. Results showed that ESEM and EUROMICRO were the two (2) top conferences that published VBSE research, each with 6 papers. We identified three (3) conferences listed in Table 11 that are no longer active: EDSER, ASWEC, and IASTED SE. EDSER used to be one of the important venues that published VBSE studies in the early 2000,

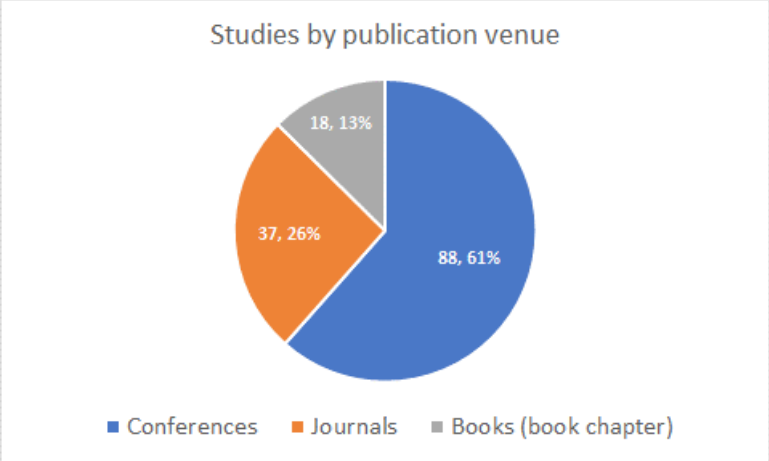

Figure 12. Number of studies by publication venues

Table 10. Classification by journals

Journal Name	Study(s)	#Studies
IEEE Software	[S7], [S19], [S22], [S23], [S42], [S61], [S102]	7
Information and Software Technology	[S91], [S108], [S135], [S138]	4
Journal of Software-Evolution and Process	[S83], [S90], [S141]	3
Software Process Improvement and Practice	[S31], [S44]	2
Software Quality Journal	[S112], [S128]	2
SIGSOFT Software Engineering Notes	[S1], [S105]	2
Requirements Engineering Journal	[S15], [S139]	2

co-located with the International Conference on Software Engineering (ICSE), a premier SE conference. The last EDSER proceedings were published in 2007. Meanwhile, the last conference for ASWEC was held in 2018, while for IASTED SE, the last conference was held in 2016.

We also analyzed the trend of publications by venue, contribution facets, and the VBSE areas (see bubble plot in Figure D5, Appendix D). Most conference papers published VBSE research in VB requirements engineering (24 studies), VB planning and control (15 studies), VB quality management (10 studies), and VB design and development (8 studies). Regarding journal publications, they focused on publishing research in two areas: VB Requirements Engineering (11 studies) and VB Planning and Control (8 studies). Finally, book chapters were published in nine different areas, with the largest number of studies in Theory of VBSE (4 studies).

Summary of key findings:
1. Most journal papers were published at IEEE Software, i.e., a magazine that targets at practitioners willing to understand applied research.
2. Only two of the 36 traditional academic journals had at least three papers published, hence suggesting a high diversity of venues for VBSE research.
3. The highest number of VBSE conference papers published in these two conferences: ESEM and EUROMICRO SEAA.

Table 11. List of publication venues (conferences)

Publication Venue	Description	Studies	Total
ESEM	Empirical Software Engineering and Measurement	[S39], [S71], [S75], [S97], [S113], [S123]	6
EUROMICRO SEAA	EUROMICRO Conference on Software Engineering and Advanced Applications	[S6], [S16], [S20], [S21], [S104], [S117]	6
EDSER	International Workshop on Economics Driven SE Research	[S69], [S111], [S132], [S134]	4
SEKE	Software Engineering and Knowledge Engineering	[S25], [S26], [S32], [S37]	4
PROFES	International Conference on Product Focused Software Development and Process Improvement	[S95], [S3], [S77], [S115]	4
ICSE	International Conference on Software Engineering	[S8], [S57], [S58]	3
ICSSP	International Conference on Software and System Process	[S98], [S73], [S35]	3
RE	Requirements Engineering Conference	[S65], [S119]	2
ICSP	International Conference on Software Process	[S74], [S93]	2
ICGSE	International Conference on Global Software Engineering	[S9], [S10]	2
ASWEC	Australasian Software Engineering Conference	[S84], [S87]	2
IASTED SE	IASTED International Conference on Software Engineering	[S29], [S38]	2
ACIS SNPD	Int'l Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing	[S12], [S64]	2
CSER	Conference on Systems Engineering Research	[S72], [S114]	2
ICEIS	International Conference on Enterprise Information Systems	[S68], [S106]	2

42 more venues with one (1) paper.

5. Discussion

5.1. Rigour and credibility issues

Prior to providing a more detailed discussion about the findings from this mapping study, with regard their implications for research and practice, we first discussed the studies' quality in relation to the publication type. Our findings clearly show that most conference papers and book chapters presented unsatisfactory or poor rigour, and acceptable credibility (illustrated as bubble plot in Figure D7, Appendix D). A few journal papers also presented poor rigour and acceptable credibility, although a higher number of journal papers present very good rigour and good credibility. With regard to journal papers, most of those showing a lower rigour and credibility failed to address criteria 6, 8 and 9; whereas regarding conference papers and book chapters, other criteria were also not addressed. The fact that

many conference papers, and even some book chapters, provided only proof of concept examples contributed significantly to their quality being assessed as lacking. All the recently published VBSE papers were empirical studies, using industrial data to assess their proposals. Many, however, are not formal experiments; therefore we believe that criterion 6, which is part of rigour, could be revisited so to also cater for other types of empirical investigations. There were also some conference papers that showed Unsatisfactory quality for Relevance. These were the 20 studies that did not even provide an example to what they were proposing. The only aspect that showed positive results for all types of publications was reporting. However, there were still 32 conference papers that had good reporting. These results, in our view, send a strong message to the VBSE community about the need to increase the rigour and credibility of VBSE studies. Conference papers also need to add additional care on the quality of their reporting and the relevance of their findings for research and/or practice.

5.2. Practical and research implications

In this Section we discussed the implications of this mapping study findings for research and practice. Except for the discussion on quality of studies (presented above), we organize the discussion based on the RQ's topic:

Value Definitions. We found that the term "value" has not been clearly or explicitly described in many VBSE studies, except for the 10% of the primary studies. VBSE authors generally regarded "value" from the perspective of worth, or utility, and not solely on economic or monetary value. Although most of the studies refer to the VBSE defined in [1] and/or [2] when describing the context of value presented in their study, the multi-dimensional perspectives of value make it difficult or challenging to measure value. Misalignment of stakeholder interest, for instance, could potentially negatively affect value because value should be measured at organizational level and therefore must be agreed upon by principle stakeholders [S47]. Hence it is important for the practitioners to do proper elicitation and reconciliation of stakeholder value propositions to avoid conflicts. Further, the difficulty for practitioners to deal with varying notions of value concept also hinders the development of software system and its features. Implication for research would be to direct research efforts towards collaborating with practitioners in developing tool support by incorporating certain specified values in software system development. This is also highlighted by Shahin et al. [32] in their study on operationalizing human values in software engineering.

Another effort that could be taken is to develop relevant measures or metrics that can be used in practice for validating values operationalization. For example, [33] developed a systematic method based on Real Option approach to manage the high level of uncertainties in requirements decision as well as to manage Technical Debts in requirements engineering. Tsilionis et al. [34] proposed a conceptual framework called "Strategic Agile Model Driven IT Governance" to ensure evaluation of value from the strategic to management level can be performed. They specifically consider three different types of value (strategic, stakeholder and user value) that could be impacted by the development or adoption of new technologies particularly in a highly dynamic business context. The complexity of measuring value is also due to the understanding that value could go beyond the monetary or utility function, e.g., value as personal attitude or beliefs, politics, culture, emotion, etc., as reported in [S139] and [S15]. [S139] developed a value-based requirements engineering method to assist the elicitation of stakeholder's value and motivation that are related to socio-political

issues in software development including the stakeholders' potential emotional reaction to system change. From our findings, we observed that measuring human related values has not received adequate attention as more research efforts focused on the utility and/or economic value of a software product or services. Ignoring human values in software development might result in user dissatisfaction and negative socio-economic impact as highlighted by [35]. Their findings showed that only a small proportion of SE research (in SE top-tier venues) directly consider human values. They mentioned, "Whilst some values (such as privacy, security, and accessibility) are well embedded in SE methods, others (such as integrity, compassion, and social justice) have received less attention". Hence, future research may consider to integrate human values in software development.

SE Principles and Practices. We identified that VB Requirements Engineering as the significant area constantly being researched since 2003. A total of 37 studies focused on various topics related to integrating value perspectives in requirements engineering, particularly looking at requirements prioritization, and contributing towards Methods, Processes, Models and Techniques. The findings indicate that VBSE research mainly focused on the early phases of software engineering (i.e., requirements engineering and software planning/control). This is because there is a lot of interest in capturing stakeholders value proposition that mostly happened at the early stage of software development project (e.g., requirements elicitation) particularly in determining the features or functionalities that should be prioritized as well as identifying the "realized value" or benefits from the software product or services [36]. Not much is known on how values can be incorporated in SE practices to analyze, prioritize and mitigate risks that occur in software project (VB Risk Management). Similarly, how VBSE can stimulate stakeholders to achieve more compatibility and improvement in terms of participation in decision making, development of shared goals and mutual trust are some new areas that can be studied, i.e., VB People Management.

Other than the VBSE domain, value consideration in software development has also been the topic of interest particularly in agile and Lean software development research (e.g., [37, 38]). According to Lane et al. [37], Lean refers to a broader concept that considers software development from the overall business perspective concentrating on the customer-defined value and waste reduction initiatives. It is interesting to note that while studies in Lean consider value end-to-end, findings from our mapping study indicate that VBSE studies focus mainly on the early phases of software development.

Our findings also indicated that since 2017 the number of publications that considered value aspects as per the value-based principles defined by [1] has declined. This might be due to the changes in the value-based research landscape, where the value concepts have been taken from different dimension since the introduction of "Value-First SE" by Ferrario et al. [39]. Value-First SE specifically uses human values as their reference framework for decision making in each software development stages. Undeniably, the emergence of unethical incidents such as the Facebook-Cambridge Analytica scandal [40] has raised the concern to embed the principles of human values in SE decision making process. Consequently, more publications on this arena have appeared, such as in Winter et al. [41], Whittle et al. [42], Ferrario et al. [43], Hussain et al. [44].

Research Topics in VBSE. We found that most of the research topics of VBSE studies fall within the area of software requirements. This finding is expected, given the large number of studies available under the VB requirements engineering domain. The fact that most VBSE studies appear under the umbrella of Requirements Engineering is inline with the results from a mapping study by [45] that showed an increasing number of SE

taxonomies in the Requirements Knowledge Area. They observed that there is a rising trends in publishing as well as utilizing SE taxonomies in recent years, particularly in the area of software requirements. Taxonomies in SE have been utilized to better structure the SE body of knowledge based on a systematic classification scheme [45].

We also observed on the low number of studies under the software maintenance topic. This is interesting as we noticed that similar findings appeared in a systematic review by [25]. They conducted a review on global software engineering, also using SWEBOK to categorize the research topics. They mentioned, "Even more notably there were no studies particularly addressing the SWEBOK knowledge areas of software construction, maintenance and configuration management, and hence these areas have been skipped in the figure." (p. 103). We highlighted as a research gap, the need to look into the areas related to software maintenance, software construction, and SE professional practice as these are recognized as important knowledge areas in software engineering according to SWEBOK [24].

Research Methods used in VBSE. Our findings indicate that case study research has been employed in most of the VBSE studies. This is not suprising, given the findings from a mapping study by [46] that showed a large number of methodological support (e.g., guidelines, supporting instruments) exists to assist researcher in performing case-study research. Their study provides a catalog of research guidelines, assessment instruments, and knowledge organization system for researcher to conduct and evaluate empirical research in SE. Molléri et al. [46] also asserted that case study methodology "is well suited for many SE research topics, as it addresses a contemporary case in depth. It aims to understand the particular case and create the basis for further research on the topic" (p. 123). We observed from this mapping study quite a large number of VBSE studies that actually did not provide empirical findings (either through validation or evaluation), hence limiting the opportunity to compare the proposed solutions. Implication for research would be to suggest SE researcher to perform necessary validation (at least) or evaluation (in real setting) on their proposed solution. Such initiatives would help increase the quality of the proposed solutions and provide better support to industrial practitioners.

Research Types in VBSE. We found that the most common research type in VBSE is solution proposal (comprised 61% of the studies), and almost half of these studies did not perform any empirical validation or evaluation. Therefore, further validation and evaluation of such proposals could be a research gap. There is also a lack of evaluation using experimental method, particularly in industrial setting. This might be due to the difficulties to arrange and conduct the experiment involving practitioners or real users. Experimental results are deemed important to enable practitioners to evaluate the proposed technique or solution, and to determine the claims made about a particular proposed solution [47]. Practitioners' commitment to participate in the evaluation studies is crucial to ensure success of a study.

Contribution Facets. VBSE research presents their contributions mostly in terms of method. Although there is a high number of a study contributed towards methods, as well as processes and models, the contributions actually appear in the form of solution proposals, which are yet to be evaluated. One notable research gap would be to develop support tools that can be used to utilize the proposed method, process, or model, which would further enable practical use of the solution, and to develop relevant metrics for evaluation or measurement purposes. Implications for practice: There have been several processes, models, and methods within VBSE studies, which can be beneficial to practitioners or

organization that currently employ VBSE principles and practices, or that are willing to use them.

Publication Venue. Our findings indicate that VBSE studies were published in various venues, mainly conferences. This is probably due to a shorter timeframe to publish in conference proceedings, when compared to journals and book chapters. Some conferences are no longer active (e.g., EDSER, IASTED SE), however premier conferences such as ESEM are EUROMICRO SEAA are still available and published VBSE related research.

5.3. Threats to validity

We have used the guidelines for reporting threats to validity for secondary studies in SE by [48]. The discussion is arranged according to the following issues: i) need for the mapping study; ii) study selection, iii) data; and iv) research.

5.3.1. Threats to validity relating to the need to conduct this mapping study

Prior to carrying out this mapping study we searched on online databases (e.g., Scopus) so to check whether there was already a mapping study or systematic literature review covering the entire field of VBSE. None were found. We initiated this work in 2016, and the second author was already aware of the fully refereed related studies that were described in Section 2.3 – [11, 12] and the grey literature [14]. The field of VBSE is an important area to SE, in particular in light of many organisations that work within a market-driven context in which different stakeholders participate in many of the decisions relating to software/software-intensive products (e.g., [4, 11]). Therefore, it was clear that the mapping study detailed herein would make a clear research contribution to VBSE.

5.3.2. Threats to validity relating to study selection

With regard to the search string and the strategy employed in this study, we used numerous synonyms around the terms value-based software engineering. There was a sub-string (``economics based'' OR ``decision making'' OR economics OR ``software project'') that was also added to our search string because it was anonymously and strongly suggested by researchers in the VBSE domain. Perhaps the final string used was quite complex; however, all the important terms were included, using several combinations of OR and AND. We also had to make a pragmatic decision in relation to the cut-off date, as initial searches showed that there would be a large number of articles to screen through, and indeed the screening, extraction of data, synthesis of results, interpretation, and writing-up has taken more than 18 months to finalise.

We screened through 6,536 titles and abstracts, which, despite the length or time needed for screening, provided us with confidence about retrieving a significant and representative sample of studies in VBSE. As we only included studies written in English, we cannot argue that our mapping study covers all studies in VBSE. Furthermore, we conducted two phases of search, which included electronic search using online databases and snowballing (backward) search. Based on the manual filtering of 3273 references from 126 primary studies, the snowballing helped discover another 17 studies. We believe we have included herein studies that represent the VBSE research population given the multi-phases search and that we employed inclusion criteria referring to the definition of VBSE. In order to validate the coverage of electronic search process, we manually checked whether the primary

studies we already knew about were retrievable from the online databases and we managed to retrieve the studies from the expected databases. All the databases employed only included fully-refereed papers; in this way we mitigated the threat of grey literature. We used a tool – Parsifal, to support most part of the study selection. This tool automatically manages duplicates, and helps with documenting the reasons for including/excluding a study.

There were numerous joint meetings to discuss the papers being screened, and the participation of the second author, who was the one with more experience in VBSE, in meetings and also the screening of titles and abstracts. This was done in order to minimize threats related to interpretive validity. Inaccuracy in data extraction and classification of studies were minimized when two researchers independently extract the data and the results were reviewed during a joint meeting. Throughout the mapping study process, several meetings (at least 8) were held and attended by the authors to discuss issues related to study selection, data extraction and classification. Each joint meeting lasts between one to four hours. In one of the joint meetings, we went through the full text of 60 studies in order to validate our selection. Any discrepancies were discussed and resolved during the joint meeting. All the authors carried out searches, and the selection of studies was checked by at least one other author, so to minimize possible biases, such as less familiarity with the VBSE research area.

5.3.3. Threats to validity relating to data

A possible threat relates to the extraction of data from all 143 primary studies, which was done by the first and the third author for all RQs, except for RQ2. However, the data extraction form and also a sample extraction were discussed in joints meetings between the first three authors, so to ensure that any possible ambiguities were solved. Some of the papers that were included were known to the second author; therefore this was also used as an additional safeguard to validate a sample of the data extracted. As for the data extracted to answer RQ2, this was done solely by the second author, who is an experienced and seasoned researcher in empirical software engineering. Furthermore, the classification used to measure their quality was chosen exactly because of its clarity. This was important so to minimise any subjectivity while extracting the data.

5.3.4. Threats to validity relating to research

Here there are two main validity threats. The first one relates to the experience of the authors. The second and fifth authors are very experienced researchers in empirical software engineering; the fifth author was the leading author in papers [15, 23], which are guidelines for conducting mapping studies in software engineering. The first author is also an experienced researcher in empirical software engineering. This mapping study also had a detailed protocol, which was followed rigorously. Furthermore, all the decisions were always discussed amongst the team, so enhancing the validity of the process that was undertaken. With regard to research generalisability, this mapping study's results are based on a sample of studies written in English, which were retrieved using a complex but wide search string. We screened through more than six thousand titles and abstracts, thus we believe that the results we present here are generalisable to the population of VBSE studies published in English.

6. Conclusions

This paper reviews VBSE studies published since 2003 with the aim to support SE community including researchers and practitioners through a collection and systematic classification of VBSE studies. We extracted value definitions and quality of studies (quality, rigour, credibility, and reporting). We classified the primary studies according to the VBSE agenda's principles and practices, research topic, research method, research type, contribution facet, and the publication venue. In this review we included 143 studies that fulfill our selection criteria and relevant to answer the research questions.

Our results showed that the term "value" has not been clearly defined in many VBSE studies, but most studies have cited either the seminal paper by [1] or the VBSE edited book [2]. In terms of quality of studies, most studies have presented very good qualiy of reporting and relevance; however, the results for studies' rigour were mixed, with the largest number of studies presenting poor rigour, followed by very good and unsatisfactory rigour. Finally, credibility was assessed as acceptable for most studies, followed by good.

The results showed that VB Requirements Engineering (37 studies, 26%) and VB Planning and Control (25 studies, 16%) were the two principles and practices mostly researched in VBSE literature, whereas VB Risk Management, VB People Management, and Other were the least researched (3% respectively). Studies in VB Requirements Engineering mostly focused on proposing new methods, processes and techniques for prioritizing requirements and mechanisms to elicit and reconcile stakeholder's value propositions.

When classified according to the SWEBOK Knowledge Area, we identified that many VBSE studies fall under the Software Requirements (34 studies, 24%) and the SE Management (30 studies, 21%) areas. In terms of research methods used by the included studies, 55 studies (38%) used case-study methodology, hence it appears to be the most common method employed. Other methods used were surveys, experiments, action research, prototyping, literature review, quantitative analysis, simulation, and mixed methods. A total of 12 studies (8%) did not declare findings, and 23 studies (16%) did not report empirical findings. While research conducted in industrial and organizational context commonly applied case study, mixed-method, and survey methodology, formal experiments are mostly conducted in academic setting

There are a small number of evaluation studies available in the VBSE literature (20 studies, 14%), and many studies either presented solution proposal (42 studies, 29%), or solution together with the validation (39 studies, 27%). Research methods used for empirical evaluation studies are mainly survey and case study. In relation to the contribution facets, most contributions were provided as methods (32 studies, 22%) and processes (31 studies, 22%), while there were very few studies that proposed metrics and tools (2 studies, 1%).

Most studies (88 studies, 61%) were conference papers presented in 57 different conference venues where ESEM, EUROMICRO SEAA, EDSER, SEKE and PROFES provided the highest number of VBSE papers. The remaining studies comprised journal papers (37 studies, 26%) and book chapters (18 studies, 13%). As part of our future work we seek to investigate how the value-based decision-making process could be influenced by the stakeholders' personality. This could potentially help address the lack of research in VB People Management in the effort to improving stakeholders' decision-making process.

Acknowledgments

This research has been carried out within the FiDiPro VALUE project number 40150/14, which is funded by Tekes (the Finnish Funding Agency for Technology and Innovation).

References

[1] B. Boehm, "Value-based software engineering: reinventing," *ACM SIGSOFT Software Engineering Notes*, Vol. 28, No. 2, Mar. 2003, p. 3.

[2] S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds., *Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006.

[3] M.W. Dictionary, "Definition of VALUE," 2022, accessed April 2022. [Online]. https://www.merriam-webster.com/dictionary/value

[4] E. Mendes, B. Turhan, P. Rodríguez, and V. Freitas, "Estimating the Value of Decisions Relating to Managing and Developing Software-intensive Products and Projects," in *Proceedings of the 11th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1–4.

[5] N. Kukreja, B. Boehm, S.S. Payyavula, and S. Padmanabhuni, "Selecting an appropriate framework for value-based requirements prioritization," in *2012 20th IEEE International Requirements Engineering Conference (RE)*, 2012, pp. 303–308.

[6] D. Port, J. Wilf, M. Diep, C. Seaman, and M. Feather, "Developing a value-based methodology for satisfying NASA software assurance requirements," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Jan. 2016, pp. 5642–5651.

[7] T. Dybå and T. Dingsøyr, "Strength of evidence in systematic reviews in software engineering," in *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 178–187.

[8] J. Favaro, "When the pursuit of quality destroys value [software development]," *IEEE Software*, Vol. 13, No. 3, 1996, pp. 93–95.

[9] B.W. Boehm, *Value-Based Software Engineering: Overview and Agenda*. Berlin, Heidelberg: Springer, 2006, Ch. 1, pp. 3–14.

[10] B.W. Boehm and A. Jain, *An Initial Theory of Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006, Ch. 2, pp. 15–37.

[11] M. Khurum, T. Gorschek, and M. Wilson, "The software value map – An exhaustive collection of value aspects for the development of software intensive products," *Journal of Software: Evolution and Process*, Vol. 25, No. 7, 2013, pp. 711–741.

[12] M.Z. Khan and M.N.A. Khan, "A Review of Value Based Software Engineering and its Impacts," *International Journal of Advanced Science and Technology*, Vol. 75, 2015, pp. 33–42.

[13] N. Salleh, F. Mendes, and E. Mendes, "A Systematic Mapping Study of Value-Based Software Engineering," in *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Aug. 2019, pp. 404–411.

[14] N. Jan and M. Ibrar, *Systematic Mapping of Value-based Software Engineering – A Systematic Review of Value-based Requirements Engineering*, Ph.D. dissertation, Blekinge Institute of Technology, Sweden, 2010. [Online]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A832388&dswid=6160

[15] K. Petersen, R. Feldt, M. Shahid, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, 2008, pp. 1–10.

[16] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, Tech. Rep., 2007.

[17] B. Kitchenham, L. Madeyski, and D. Budgen, "Segress: Software engineering guidelines for reporting secondary studies," *IEEE Transactions on Software Engineering*, Vol. 49, No. 3, 2023, pp. 1273–1298.

[18] S. Jalali and C. Wohlin, "Systematic literature studies: Database searches vs. backward snowballing," in *Proceedings of the 2012 ACM-IEEE international symposium on empirical software engineering and measurement*. IEEE, 2012, pp. 29–38.

[19] D. Maplesden, E. Tempero, J. Hosking, and J. Grundy, "Performance analysis for object-oriented software: A systematic mapping," *IEEE Transactions on Software Engineering*, Vol. 41, No. 7, 2015, pp. 691–710.

[20] T. Dyba, T. Dingsoyr, and G. Hanssen, "Applying systematic reviews to diverse study types: An experience report," in *First International Symposium on Empirical Software Engineering and Measurement (ESEM2007)*. IEEE, 2007, pp. 225–234.

[21] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, Vol. 55, No. 12, 2013, pp. 2049–2075.

[22] H. Zhang, M.A. Babar, and P. Tell, "Identifying relevant studies in software engineering," *Information and Software Technology*, Vol. 53, No. 6, 2011, pp. 625–637.

[23] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, Vol. 64, 2015, pp. 1–18.

[24] P. Bourque and R. Fairley, "Society IC (2014) guide to the software engineering body of knowledge (SWEBOK®): Version 3.0."

[25] D. Šmite, C. Wohlin, T. Gorschek, and R. Feldt, "Empirical evidence in global software engineering: A systematic review," *Empirical Software Engineering*, Vol. 15, No. 1, 2010, pp. 91–118.

[26] S. Easterbrook, J. Singer, M.A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," in *Guide to Advanced Empirical Software Engineering*. Springer, 2008, pp. 285–311.

[27] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requirements engineering*, Vol. 11, No. 1, 2006, pp. 102–107.

[28] S.H. Schwartz, "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," in *Advances in Experimental Social Psychology*. Elsevier, 1992, Vol. 25, pp. 1–65.

[29] J.S. Molléri, K. Petersen, and E. Mendes, "Towards understanding the relation between citations and research quality in software engineering studies," *Scientometrics*, Vol. 117, No. 3, 2018, pp. 1453–1478.

[30] T. Dingsøyr and C. Lassenius, "Emerging themes in agile software development," *Information and Software Technology*, Vol. 77, 2016, pp. 56–60.

[31] A. Shahrokni and R. Feldt, "A systematic review of software robustness," *Information and Software Technology*, Vol. 55, No. 1, 2013, pp. 1–17.

[32] M. Shahin, W. Hussain, A. Nurwidyantoro, H. Perera, R. Shams et al., "Operationalizing human values in software engineering: A survey," *IEEE Access*, Vol. 10, 2022, pp. 75 269–75 295.

[33] Z.S.H. Abad and G. Ruhe, "Using real options to manage technical debt in requirements engineering," in *2015 IEEE 23rd International Requirements Engineering Conference (RE)*. IEEE, 2015, pp. 230–235.

[34] K. Tsilionis and Y. Wautelet, "A model-driven framework to support strategic agility: Value-added perspective," *Information and Software Technology*, Vol. 141, 2022.

[35] H. Perera, W. Hussain, J. Whittle, A. Nurwidyantoro, D. Mougouei et al., "A study on the prevalence of human values in software engineering publications, 2015–2018," in *IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*. IEEE, 2020, pp. 409–420.

[36] T. Gilb and L. Brodie, "What's fundamentally wrong? improving our approach towards capturing value in requirements specification," in *INCOSE International Symposium*, Vol. 22, No. 1. Wiley Online Library, 2012, pp. 926–939.

[37] M. Lane, B. Fitzgerald, and P. Ågerfalk, "Identifying lean software development values," 2012.
[38] P. Middleton, "Lean software development: Two case studies," *Software Quality Journal*, Vol. 9, No. 4, 2001, pp. 241–252.
[39] M.A. Ferrario, W. Simm, S. Forshaw, A. Gradinar, M.T. Smith et al., "Values-first SE: Research principles in practice," in *IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C)*, 2016, pp. 553–562.
[40] J.C. Wong, "The Cambridge Analytica scandal change the world – But it didn't change Facebook," 2019, accessed April 2022. [Online]. https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook
[41] E. Winter, S. Forshaw, L. Hunt, and M. Ferrario, "Advancing the study of human values in software engineering," in *IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, 2019, pp. 19–26.
[42] J. Whittle, M.A. Ferrario, W. Simm, and W. Hussain, "A case for human values in software engineering," *IEEE Software*, Vol. 38, No. 1, 2021, pp. 106–113.
[43] M.A. Ferrario and E. Winter, "Applying human values theory to software engineering practice: Lessons and implications," *IEEE Transactions on Software Engineering*, 2022, p. 1.
[44] W. Hussain, H. Perera, J. Whittle, A. Nurwidyantoro, R. Hoda et al., "Human values in software engineering: Contrasting case studies of practice," *IEEE Transactions on Software Engineering*, Vol. 48, No. 5, 2022, pp. 1818–1833.
[45] M. Usman, R. Britto, J. Börstler, and E. Mendes, "Taxonomies in software engineering: A systematic mapping study and a revised taxonomy development method," *Information and Software Technology*, Vol. 85, 2017, pp. 43–59.
[46] J.S. Molléri, K. Petersen, and E. Mendes, "CERSE-catalog for empirical research in software engineering: A systematic mapping study," *Information and Software Technology*, Vol. 105, 2019, pp. 117–149.
[47] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell et al., *Experimentation in software engineering*. Springer Science and Business Media, 2012.
[48] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," *Information and Software Technology*, Vol. 106, 2019, pp. 201–230.
[49] R.L. Keeney, H. Raiffa, and D.W. Rajala, "Decisions with multiple objectives: Preferences and value trade-offs," *IEEE transactions on Systems, man, and cybernetics*, Vol. 9, No. 7, 1979, pp. 403–403.
[50] P. Ojala, "Value of project management: a case study," *WSEAS Transactions on Information Science and Applications*, Vol. 6, No. 3, 2009, p. 2009.
[51] I. Ramos and D.M. Berry, "Is emotion relevant to requirements engineering?" *Requirements Engineering*, Vol. 10, No. 3, 2005, pp. 238–242.

Appendix A. Primary studies

List of selected studies

[S1] B. Boehm, "Value-based software engineering," *ACM SIGSOFT Software Engineering Notes*, Vol. 28, No. 2, Mar. 2003, p. 4.

[S2] T. Dingsøyr, "Value-based knowledge management: The contribution of group processes," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 309–325.

[S3] C. Fernández, D. López, A. Yagüe, and J. Garbajosa, "Towards estimating the value of an idea," in *Proceedings of the 12th International Conference on Product Focused Software Development and Process Improvement*, Profes '11. New York, NY, USA: Association for Computing Machinery, Jun. 2011, pp. 62–67.

[S4] E. Polis, "Value and viability considerations in information systems development," in *Proceedings of the 2011 conference on Databases and Information Systems VI: Selected Papers from the Ninth International Baltic Conference, DB&IS 2010*. NLD: IOS Press, Aug. 2011, pp. 257–270.

[S5] N.A. Zakaria, S. Ibrahim, and M.N. Mahrin, "A proposed value-based software process tailoring framework," in *9th Malaysian Software Engineering Conference (MySEC)*, Dec. 2015, pp. 149–153.

[S6] R. Ramler, T. Kopetzky, and W. Platz, "Value-based coverage measurement in requirements-based testing: Lessons learned from an approach implemented in the TOSCA testsuite," in *38th Euromicro Conference on Software Engineering and Advanced Applications*, Sep. 2012, pp. 363–366.

[S7] S. Barney, G. Hu, A. Aurum, and C. Wohlin, "Creating software product value in China," *IEEE Software*, Vol. 26, No. 4, Jul. 2009, pp. 84–90.

[S8] N. Kukreja, "Decision theoretic requirements prioritization A two-step approach for sliding towards value realization," in *35th International Conference on Software Engineering (ICSE)*, May 2013, pp. 1465–1467.

[S9] R. Bavani, "Global software engineering: Challenges in customer value creation," in *5th IEEE International Conference on Global Software Engineering*, Aug. 2010, pp. 119–122.

[S10] S. Barney, C. Wohlin, P. Chatzipetrou, and L. Angelis, "Offshore insourcing: A case study on software quality alignment," in *IEEE Sixth International Conference on Global Software Engineering*, Aug. 2011, pp. 146–155.

[S11] A. Murtazaev, S. Kang, J. Baik, and J. Lee, "An approach to defining a value-based software development process," in *IEEE/ACIS 9th International Conference on Computer and Information Science*, Aug. 2010, pp. 690–695.

[S12] C.K. Kim, D.H. Lee, I.Y. Ko, and J. Baik, "A lightweight value-based software architecture evaluation," in *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, Vol. 2. IEEE, 2007, pp. 646–649.

[S13] M. Ramzan, M.A. Jaffar, M.A. Iqbal, S. Anwar, and A.A. Shahid, "Value based fuzzy requirement prioritization and its evaluation framework," in *Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*. IEEE, 2009, pp. 1464–1468.

[S14] F. Sher, D.N.A. Jawawi, R. Mohamad, and M.I. Babar, "Multi-aspects based requirements priortization technique for value-based software developments," in *International Conference on Emerging Technologies (ICET)*, Dec. 2014, pp. 1–6.

[S15] J. Zdravkovic, E.O. Svee, and C. Giannoulis, "Capturing consumer preferences as requirements for software product lines," *Requirements Engineering*, Vol. 20, No. 1, Mar. 2015, pp. 71–90.

[S16] D. Wahyudin, A. Schatten, D. Winkler, and S. Biffl, "Aspects of software quality assurance in open source software projects: Two case studies from Apache Project," in *33rd EUROMICRO Conference on Software Engineering and Advanced Applications*, Aug. 2007, pp. 229–236.

[S17] R. Yin, H. Hu, J. Ge, and J. Lu, "Quantitative analysis of value-based software processes using decision-based stochastic object Petri-Nets," in *14th Asia-Pacific Software Engineering Conference (APSEC'07)*, Dec. 2007, pp. 526–533.

[S18] S.W. Lim, T. Lee, S. Kim, and H.P. In, "The value gap model: Value-based requirements elicitation," in *7th IEEE International Conference on Computer and Information Technology (CIT 2007)*, Oct. 2007, pp. 885–890.

[S19] J. Azar, R.K. Smith, and D. Cordes, "Value-oriented requirements prioritization in a small development organization," *IEEE Software*, Vol. 24, No. 1, Jan. 2007, pp. 32–37.

[S20] K.W. Wagner and W. Durr, "A five-step method for value-based planning and monitoring of systems engineering projects," in *32nd EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO'06)*, Aug. 2006, pp. 282–290.

[S21] M. Heindl, F. Reinisch, S. Biffl, and A. Egyed, "Value-based selection of requirements engineering tool support," in *32nd EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO'06)*, Aug. 2006, pp. 266–273.

[S22] L. Huang and B. Boehm, "How much software quality investment is enough: A value-based approach," *IEEE Software*, Vol. 23, No. 5, Sep. 2006, pp. 88–95.

[S23] B. Boehm, L. Huang, A. Jain, and R. Madachy, "The ROI of software dependability: The iDAVE model," *IEEE Software*, Vol. 21, No. 3, May 2004, pp. 54–61.

[S24] L. Brodie and M. Woodman, "Prioritization of stakeholder value using metrics," in *Evaluation of Novel Approaches to Software Engineering*, Communications in Computer and Information Science, L.A. Maciaszek and P. Loucopoulos, Eds. Berlin, Heidelberg: Springer, 2011, pp. 74–88.

[S25] D. Zhang, "Taming inconsistency in value-based software development," in *Proceedings of the Twenty-First International Conference on Software Engineering and Knowledge Engineering*, Boston, Massachusetts, Jul. 2009, pp. 450–455.

[S26] D. Zhang, "Machine learning and value-based software engineering: A research agenda," in *The 20th International Conference on Software Engineering and Knowledge Engineering*, San Francisco Bay, USA, 2008, pp. 285–290.

[S27] J.K. Balikuddembe and A. Bagula, "Aligning the software project selection process with the business strategy: A pilot study," in *Advances in Software Engineering*, Communications in Computer and Information Science, D. Ślęzak, T.H. Kim, A. Kiumi, T. Jiang, J. Verner et al., Eds. Berlin, Heidelberg: Springer, 2009, pp. 237–244.

[S28] G. Hoff, A. Fruhling, and K. Ward, "Requirement prioritization decision factors for agile development environments," in *AMCIS 2008 Proceedings*, Jan. 2008.

[S29] J. Samad, N. Ikram, and M. Usman, "VRRM: A value-based requirements' risk management process," in *Proceedings of the IASTED International Conference on Software Engineering*, SE '08. USA: ACTA Press, Feb. 2008, pp. 184–191.

[S30] S. Barney, A. Aurum, and C. Wohlin, "A product management challenge: Creating software product value through requirements selection," *Journal of Systems Architecture*, Vol. 54, No. 6, Jun. 2008, pp. 576–593.

[S31] B. Boehm and A. Jain, "Developing a process framework using principles of value-based software engineering: Research sections," *Software Process: Improvement and Practice*, Vol. 12, No. 5, Sep. 2007, pp. 377–385.

[S32] S. Ziemer, P.R.F. Sampaio, and T. Stalhane, "A decision modelling approach for analysing requirements configuration trade-offs in time-constrained web application development," in *18th International Conference on Software Engineering and Knowledge Engineering, SEKE*, 2006, pp. 144–149.

[S33] H. Alahyari, R. Berntsson Svensson, and T. Gorschek, "A study of value in agile software development organizations," *Journal of Systems and Software*, Vol. 125, Mar. 2017, pp. 271–288.

[S34] E. Mendes, V. Freitas, M. Perkusich, J. Nunes, F. Ramos et al., "Using Bayesian network to estimate the value of decisions within the context of value-based software engineering: A multiple case study," *International journal of software engineering and knowledge engineering*, Vol. 29, No. 11–12, 2019, pp. 1629–1671.

[S35] N. Kukreja and B. Boehm, "Integrating collaborative requirements negotiation and prioritization processes: A match made in heaven," in *Proceedings of the 2013 International Conference on Software and System Process*, ICSSP 2013. New York, NY, USA: Association for Computing Machinery, May 2013, pp. 141–145.

[S36] D. Lettner, D. Thaller, M. Vierhauser, R. Rabiser, P. Grünbacher et al., "Supporting business calculations in a product line engineering tool suite," in *Proceedings of the 15th International Software Product Line Conference, Volume 2*, SPLC '11. New York, NY, USA: Association for Computing Machinery, Aug. 2011, pp. 1–4.

[S37] D. Zhang, "Capturing antagonistic stakeholder value propositions in value-based software development," in *Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering (SEKE'2010)*, Redwood City, San Francisco Bay, CA, USA, 2010.

[S38] N. Ahmad, M. Usman, and N. Ikram, "Value-based software architecture knowledge management tool," in *Proceedings of the IASTED International Conference on Software Engineering*. Innsbruck, Austria: ACTAPRESS, 2010.

[S39] V. Mandić, V. Basili, L. Harjumaa, M. Oivo, and J. Markkula, "Utilizing GQM + Strategies for business value analysis: An approach for evaluating business goals," in *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '10. New York, NY, USA: Association for Computing Machinery, Sep. 2010.

[S40] H. Sneed and S. Huang, "Value-driven software maintenance," *International Journal of Computers and Applications*, Vol. 32, No. 2, Jan. 2010.

[S41] L. Huang, "A value-based process for achieving software dependability," in *Unifying the Software Process Spectrum*, Lecture Notes in Computer Science, M. Li, B. Boehm, and L.J. Osterweil, Eds. Berlin, Heidelberg: Springer, 2006, pp. 108–121.

[S42] Y. Yang, J. Bhuta, B. Boehm, and D. Port, "Value-based processes for COTS-based applications," *IEEE Software*, Vol. 22, No. 4, Jul. 2005, pp. 54–62.

[S43] G. Murtaza, N. Ikram, and A. Basit, "A framework for eliciting value proposition from stakeholders," *WSEAS Transactions on Computers*, Vol. 9, No. 6, Jun. 2010, pp. 557–572.

[S44] S. Biffl, D. Winkler, R. Höhn, and H. Wetzel, "Software process improvement in Europe: potential of the new V-modell XT and research issues," *Software Process: Improvement and Practice*, Vol. 11, No. 3, 2006, pp. 229–238.

[S45] A. Fruhling and G.J. de Vreede, "Collaborative usability testing to facilitate stakeholder involvement," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 201–223.

[S46] R. Ramler, S. Biffl, and P. Grünbacher, "Value-based management of software testing," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 225–244.

[S47] H. Erdogmus, J. Favaro, and M. Halling, "Valuation of software initiatives under uncertainty: Concepts, issues, and techniques," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 39–66.

[S48] B.W. Boehm and A. Jain, "An initial theory of value-based software engineering," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 15–37.

[S49] B.W. Boehm, "Value-based software engineering: Overview and agenda," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 3–14.

[S50] B.W. Boehm, "Value-based software engineering: Seven key elements and ethical considerations," in *Value-Based Software Engineering*, S. Biffl, A. Aurum, B. Boehm, H. Erdogmus, and P. Grünbacher, Eds. Berlin, Heidelberg: Springer, 2006, pp. 109–132.

[S51] M. Berry and A. Aurum, "Measurement and decision making," in *Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006, pp. 155–177.

[S52] A. Gachet and R. Sprague, "A context-based approach to the development of decision support systems," in *International Workshop on Context Modeling and Decision Support*, Paris, France, 2005.

[S53] L. Huang, H. Hu, J. Ge, B. Boehm, and J. Lü, "Tailor the value-based software quality achievement process to project business cases," in *Software Process Change*. Berlin, Heidelberg: Springer, 2006, pp. 56–63.

[S54] E. Mendes, B. Turhan, P. Rodríguez, and V. Freitas, "Estimating the value of decisions relating to managing and developing software-intensive products and projects," in *Proceedings of the 11th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1–4.

[S55] T.J. Latha and L. Suganthi, "An empirical study on creating software product value in India – An analytic hierarchy process approach," *International Journal of Business Information Systems*, Vol. 18, No. 1, Dec. 2015, pp. 26–43.

[S56] Y. Han, D.h. Lee, B. Choi, M. Hinchey, and H.P. In, "Value-driven V-model: From requirements analysis to acceptance testing," *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, 2016, pp. 1776–1785.

[S57] L. Huang, B. Boehm, H. Hu, J. Ge, J. Lü et al., "Applying the Value/Petri process to ERP software development in China," in *Proceedings of the 28th international conference on Software engineering*, ICSE '06. New York, NY, USA: Association for Computing Machinery, May 2006, pp. 502–511.

[S58] D. Falessi, R. Capilla, and G. Cantone, "A value-based approach for documenting design decisions rationale: a replicated experiment," in *Proceedings of the 3rd international workshop on Sharing and reusing architectural knowledge*, SHARK '08. New York, NY, USA: Association for Computing Machinery, May 2008, pp. 63–70.

[S59] D. Zhang, "A value-based framework for software evolutionary testing," *International Journal of Software Science and Computational Intelligence (IJSSCI)*, Vol. 3, No. 2, Apr. 2011, pp. 62–82.

[S60] S.S. Payyavula, S.S. Jahagirdar, and M. Kumar, "Application of value based requirement prioritization in a banking product implementation," in *Third International Conference on Services in Emerging Markets*, Dec. 2012, pp. 157–161.

[S61] S. Kim, H.P. In, J. Baik, R. Kazman, and K. Han, "VIRE: Sailing a blue ocean with value-innovative requirements," *IEEE Software*, Vol. 25, No. 1, Jan. 2008, pp. 80–87.

[S62] X. Zhang, G. Auriol, and C. Baron, "Understanding customer expectations for system development," in *Fifth International Conference on Software Engineering Advances*, Aug. 2010, pp. 44–49.

[S63] R.P. dos Santos, L.R. Tostes, and C.M.L. Werner, "A Brechó-EcoSys extension to support negotiation in the software ecosystems context," in *IEEE 14th International Conference on Information Reuse and Integration (IRI)*, Aug. 2013, pp. 578–585.

[S64] N. Kim, T. Lee, D. Lee, K. Lee, and H.P. In, "Customer value-based HW/SW partitioning decision in embedded systems," in *Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, Aug. 2008, pp. 257–262.

[S65] N. Kukreja, B. Boehm, S.S. Payyavula, and S. Padmanabhuni, "Selecting an appropriate framework for value-based requirements prioritization," in *20th IEEE International Requirements Engineering Conference (RE)*, Sep. 2012, pp. 303–308.

[S66] D. Cabrero, J. Garzás, and M. Piattini, "Choosing the best design strategy from requirements. A value-based approach," in *IEEE International Conference on Exploring Quantifiable IT Yields*, Mar. 2007, pp. 87–94.

[S67] A. Jain and B. Boehm, "SimVBSE: Developing a game for value-based software engineering," in *19th Conference on Software Engineering Education and Training (CSEET '06)*, Apr. 2006, pp. 103–114.

[S68] A. Itaborahy, K. Oliveira, and R. Santos, "Value-based software project management – A business perspective on software projects," in *International Conference on Enterprise Information Systems*, Jan. 2008, pp. 218–225.

[S69] A. Jain and B. Boehm, "Developing a theory of value-based software engineering," in *Proceedings of the seventh international workshop on Economics-driven software engineering research*, EDSER '05. New York, NY, USA: Association for Computing Machinery, May 2005, pp. 1–5.

[S70] B. Boehm and L.G. Huang, "Value-based software engineering: A case study," *Computer*, Vol. 36, No. 3, Mar. 2003, pp. 33–41.

[S71] K. Lee and B. Boehm, "Empirical results from an experiment on value-based review (VBR) processes," in *International Symposium on Empirical Software Engineering, 2005*, Nov. 2005, p. 10.

[S72] N. Kukreja, S.S. Payyavula, B. Boehm, and S. Padmanabhuni, "Value-based requirements prioritization: Usage experiences," *Procedia Computer Science*, Vol. 16, Jan. 2013, pp. 806–813.

[S73] Q. Li, B. Boehm, Y. Yang, and Q. Wang, "A value-based review process for prioritizing artifacts," in *Proceedings of the International Conference on Software and Systems Process*, ICSSP '11. New York, NY, USA: Association for Computing Machinery, May 2011, pp. 13–22.

[S74] R. Madachy and B. Boehm, "Assessing quality processes with ODC COQUALMO," in *Making Globally Distributed Software Development a Success Story*. Berlin, Heidelberg: Springer, 2008, pp. 198–209.

[S75] L. Huang and B. Boehm, "Determining how much software assurance is enough? A value-based approach," in *International Symposium on Empirical Software Engineering, 2005*, Nov. 2005, p. 10.

[S76] B. Boehm and A. Jain, "A value-based software process framework," in *Software Process Change*. Berlin, Heidelberg: Springer, 2006, pp. 1–10.

[S77] M. Rönkkö, C. Frühwirth, and S. Biffl, "Integrating value and utility concepts into a value decomposition model for value-based software engineering," in *Product-Focused Software Process Improvement*. Berlin, Heidelberg: Springer, 2009, pp. 362–374.

[S78] R. dos Santos, M. Silva, and C. Werner, "Brechó-VCM: A value-based approach for component markets," *International Transactions on Systems Science and Applications*, Vol. 6, No. 2/3, 2010, pp. 179–199.

[S79] S.I. Mohamed and A.M. Wahba, "Value estimation for software product management," in *IEEE International Conference on Industrial Engineering and Engineering Management*, Dec. 2008, pp. 2196–2200.

[S80] D. Falessi, G. Cantone, and P. Kruchten, "Value-based design decision rationale documentation: Principles and empirical feasibility study," in *Proceedings of the Seventh Working IEEE/IFIP Conference on Software Architecture (WICSA 2008)*, WICSA '08. USA: IEEE Computer Society, Feb. 2008, pp. 189–198.

[S81] G.S. de Aquino and S.R. de Lemos Meira, "An approach to measure value-based productivity in software projects." IEEE Computer Society, Aug. 2009, pp. 383–389.

[S82] D. Raffo, M. Mehta, D.J. Anderson, and R. Harmon, "Integrating Lean principles with value based software engineering," in *PICMET Technology Management for Global Economic Growth*, Jul. 2010, pp. 1–10.

[S83] Q. Li, Y. Yang, M. Li, Q. Wang, B.W. Boehm et al., "Improving software testing process: Feature prioritization to make winners of success-critical stakeholders," *Journal of Software: Evolution and Process*, Vol. 24, No. 7, 2012, pp. 783–801.

[S84] S. Marciuska, C. Gencel, and P. Abrahamsson, "Feature usage as a value indicator for decision making," in *23rd Australian Software Engineering Conference*, Apr. 2014, pp. 124–131.

[S85] D. Falessi, M. Becker, and G. Cantone, "Design decision rationale: Experiences and steps ahead towards systematic use," *ACM SIGSOFT Software Engineering Notes*, Vol. 31, No. 5, Sep. 2006, p. 2.

[S86] M. Heindl and S. Biffl, "A case study on value-based requirements tracing," in *Proceedings of the 10th European software engineering conference held jointly with 13th ACM SIGSOFT international symposium on Foundations of software engineering*, ESEC/FSE-13. New York, NY, USA: Association for Computing Machinery, Sep. 2005, pp. 60–69.

[S87] N.A. Zakaria, S. Ibrahim, and M.N. Mahrin, "Examining value-based factors in software development: A survey study in Malaysian public sector," in *Proceedings of the ASWEC 24th Australasian Software Engineering Conference*, ASWEC '15, Vol. II. New York, NY, USA: Association for Computing Machinery, Sep. 2015, pp. 13–17.

[S88] S. Marciuska, C. Gencel, and P. Abrahamsson, "Exploring how feature usage relates to customer perceived value: A case study in a startup company," in *Software Business. From Physical Products to Software Services and Solutions*. Berlin, Heidelberg: Springer, 2013, pp. 166–177.

[S89] A. Ivanović, P. America, and C. Snijders, "Modeling customer-centric value of system architecture investments," *Software & Systems Modeling*, Vol. 12, No. 2, May 2013, pp. 369–385.

[S90] M. Khurum, T. Gorschek, and M. Wilson, "The software value map – Aan exhaustive collection of value aspects for the development of software intensive products," *Journal of Software: Evolution and Process*, Vol. 25, No. 7, 2013, pp. 711–741.

[S91] S. Barney, V. Mohankumar, P. Chatzipetrou, A. Aurum, C. Wohlin et al., "Software quality across borders: Three case studies on company internal alignment," *Information and Software Technology*, Vol. 56, No. 1, Jan. 2014, pp. 20–38.

[S92] A. Egyed, S. Biffl, M. Heindl, and P. Grünbacher, "A value-based approach for understanding cost-benefit trade-offs during automated software traceability," in *Proceedings of the 3rd international workshop on Traceability in emerging forms of software engineering*, TEFSE '05. New York, NY, USA: Association for Computing Machinery, Nov. 2005, pp. 2–7.

[S93] Q. Li, M. Li, Y. Yang, Q. Wang, T. Tan et al., "Bridge the gap between software test process and business value: A case study," in *Trustworthy Software Development Processes*. Berlin, Heidelberg: Springer, 2009, pp. 212–223.

[S94] R. Vetschera, "Preference-based decision support in software engineering," in *Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006, pp. 67–89.

[S95] O. Castro, A. Espinoza, and A. Martínez-Martínez, "Estimating the software product value during the development process," in *Product-Focused Software Process Improvement*. Berlin, Heidelberg: Springer, 2012, pp. 74–88.

[S96] S. Maurice, G. Ruhe, O. Saliu, and A. Ngo-The, "Decision support for value-based software release planning," in *Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006, pp. 247–261.

[S97] C. Wohlin and A. Aurum, "What is important when deciding to include a software requirement in a project or release?" in *International Symposium on Empirical Software Engineering, 2005*, Nov. 2005, p. 10.

[S98] Q. Li and B. Boehm, "Improving scenario testing process by adding value-based prioritization: An industrial case study," in *Proceedings of the International Conference on Software and System Process*, ICSSP 2013. New York, NY, USA: Association for Computing Machinery, May 2013, pp. 78–87.

[S99] P. Grünbacher, S. Köszegi, and S. Biffl, "Stakeholder value proposition elicitation and reconciliation," in *Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006, pp. 133–154.

[S100] R. Madachy, B. Boehm, J. Richardson, M. Feather, and T. Menzies, "Value-based design of software V&V processes for NASA flight projects," in *AIAA SPACE Conference and Exposition*, AIAA SPACE Forum. American Institute of Aeronautics and Astronautics, Sep. 2007.

[S101] A. Aurum and C. Wohlin, "A value-based approach in requirements engineering: Explaining some of the fundamental concepts," in *Requirements Engineering: Foundation for Software Quality*. Berlin, Heidelberg: Springer, 2007, pp. 109–115.

[S102] B. Boehm and J. Bhuta, "Balancing opportunities and risks in component-based software development," *IEEE Software*, Vol. 25, No. 6, Nov. 2008, pp. 56–63.

[S103] M.I. Babar, M. Ghazali, and D.N.A. Jawawi, "Risk based decision support system for stakeholder quantification for value based software systems," *Journal of Theoretical and Applied Information Technology*, Vol. 76, No. 3, Jun. 2015, pp. 373–385.

[S104] S. Barney, A. Aurum, and C. Wohlin, "Quest for a silver bullet: Creating software product value through requirements selection," in *32nd EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO'06)*, Aug. 2006, pp. 274–281.

[S105] B. Boehm, "Value-based software engineering: Reinventing," *ACM SIGSOFT Software Engineering Notes*, Vol. 28, No. 2, Mar. 2003, p. 3.

[S106] D. Moreno, J. Garzás, and M. Piattini, "Maintenance cost of a software design: A value-based approach," in *Proceedings of the international conference on Enterprise Information System*, Jan. 2007, pp. 384–389.

[S107] G. Hu, A. Aurum, and C. Wohlin, "Adding value to software requirements: An empirical study in the Chinese software industry," in *ACIS Proceedings*, Jan. 2006.

[S108] H. Huijgens, A. van Deursen, and R. van Solingen, "The effects of perceived value and stakeholder satisfaction on software project impact," *Information and Software Technology*, Vol. 89, Sep. 2017, pp. 19–36.

[S109] H. In and D. Olson, "Requirements negotiation using multi-criteria preference analysis," *JUCS – Journal of Universal Computer Science*, Vol. 10, No. 4, Apr. 2004, pp. 306–325.

[S110] M. Ramzan, M.A. Jaffar, and A.A. Shahid, "Value assignment process (VAP): Establishing value of software through a new definition of value," in *Proceedings of the 4th International Conference on Ubiquitous Information Technologies and Applications*, Dec. 2009, pp. 1–8.

[S111] R. Madachy, "Simulation for business value and software process/product tradeoff decisions," in *Proceedings of the international workshop on Economics driven software engineering research*, EDSER '06. New York, NY, USA: Association for Computing Machinery, May 2006, pp. 25–30.

[S112] E. Mendes, P. Rodriguez, V. Freitas, S. Baker, and M.A. Atoui, "Towards improving decision making and estimating the value of decisions in value-based software engineering: the VALUE framework," *Software Quality Journal*, Vol. 26, No. 2, Jun. 2018, pp. 607–656.

[S113] D. Port, T. Bui, J. Wilf, Y. Kobayashi, and Y. Miyamoto, "What we have learned about the value of software assurance," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '14. New York, NY, USA: Association for Computing Machinery, Sep. 2014, pp. 1–8.

[S114] D. Port and J. Wilf, "The value proposition for assurance of JPL systems," *Procedia Computer Science*, Vol. 28, Jan. 2014, pp. 398–403.

[S115] Z. Racheva, M. Daneva, K. Sikkel, and L. Buglione, "Business value is not only dollars – Results from case study research on agile software projects," in *Product-Focused Software Process Improvement*. Berlin, Heidelberg: Springer, 2010, pp. 131–145.

[S116] M. Ramzan, M. Jaffar, and A.A. Shahid, "Value-based intelligent requirement prioritization (VIRP): Expert driven fuzzy logic based prioritization technique," *International Journal of Innovative Computing, Information and Control*, Vol. 7, No. 3, 2011.

[S117] V. Freitas, E. Mendes, and B. Turhan, "Providing tool-support for value-based decision-making: A usability assessment," in *42th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2016, pp. 34–41.

[S118] M. Yilmaz, R.V. O'Connor, and J. Collins, "Improving software development process through economic mechanism design," in *Systems, Software and Services Process Improvement*. Berlin, Heidelberg: Springer, 2010, pp. 177–188.

[S119] Z. Racheva, M. Daneva, K. Sikkel, A. Herrmann, and R. Wieringa, "Do we know enough about requirements prioritization in agile projects: Insights from a case study," in *18th IEEE International Requirements Engineering Conference*, Sep. 2010, pp. 147–156.

[S120] N.A. Zakaria, S. Ibrahim, and M.N. Mahrin, "An integrated approach to formulate a value-based software process tailoring framework," *Jurnal Teknologi*, Vol. 78, No. 12–3, 2016, pp. 171–180.

[S121] X. Zhu and B. Zhou, "An earned-value approach to assess and monitor software project uncertainty: A case study in software test execution," *Information Technology Journal*, Vol. 9, No. 6, Jun. 2010, pp. 1104–1114.

[S122] H.M. Chen, R. Kazman, J. Garbajosa, and E. Gonzalez, "Toward big data value engineering for innovation," in *Proceedings of the 2nd International Workshop on BIG Data Software Engineering*, BIGDSE '16. New York, NY, USA: Association for Computing Machinery, May 2016, pp. 44–50.

[S123] D. Port and J. Wilf, "The value of certifying software release readiness: An exploratory study of certification for a critical system at JPL," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, Oct. 2013, pp. 373–382.

[S124] A.K. Gupta and A. Deraman, "Algorithmic solution for effective selection of elicitation techniques," in *International Conference on Computer and Information Sciences (ICCIS)*, Apr. 2019, pp. 1–7.

[S125] D. Port and J. Wilf, "A study on the perceived value of software quality assurance at JPL," in *44th Hawaii International Conference on System Sciences*, Jan. 2011, pp. 1–10.

[S126] C. Wohlin and A. Aurum, "Criteria for selecting software requirements to create product value: An industrial empirical study," in *Value-Based Software Engineering*. Berlin, Heidelberg: Springer, 2006, pp. 179–200.

[S127] J. Favaro, "Value based management and agile methods," in *Extreme Programming and Agile Processes in Software Engineering*. Berlin, Heidelberg: Springer, 2003, pp. 16–25.

[S128] B. Wong, "Understanding stakeholder values as a means of dealing with stakeholder conflicts," *Software Quality Journal*, Vol. 13, No. 4, Dec. 2005, pp. 429–445.

[S129] C. Werner, L. Murta, A. Marinho, R. dos Santos, and M. Silva, "Towards a component and service marketplace with Brechó Library," in *Proceedings of the IADIS International Conference WWW/Internet*, Vol. 1, Nov. 2009, pp. 567–574.

[S130] A. Egyed, P. Grunbacher, M. Heindl, and S. Biffl, "Value-based requirements traceability: Lessons learned," in *15th IEEE International Requirements Engineering Conference (RE 2007)*, Oct. 2007, pp. 115–118.

[S131] M. Book, S. Grapenthin, and V. Gruhn, "Value-based migration of legacy data structures," in *Software Quality. Model-Based Approaches for Advanced Software and Systems Engineering*. Cham: Springer, 2014, pp. 115–134.

[S132] C. Scaffidi, A. Arora, S. Butler, and M. Shaw, "A value-based approach to predicting system properties from design," *ACM SIGSOFT Software Engineering Notes*, Vol. 30, No. 4, May 2005, pp. 1–5.

[S133] D. Sobhy, R. Bahsoon, L. Minku, and R. Kazman, "Diversifying software architecture for sustainability: A value-based perspective," in *Software Architecture*. Cham: Springer, 2016, pp. 55–63.

[S134] V. Poladian, S. Butler, M. Shaw, and D. Garlan, "Time is not money: The case for multi-dimensional accounting in value-based software engineering," in *Fifth Workshop on Economics-Driven Software Engineering Research (EDSER-5)*, May 2003.

[S135] A.M. Pitangueira, P. Tonella, A. Susi, R.S.P. Maciel, and M. Barros, "Minimizing the stakeholder dissatisfaction risk in requirement selection for next release planning," *Information and Software Technology*, Vol. 87, Jul. 2017, pp. 104–118.

[S136] V.J.A.T. de Melo França, R. Balancieri, G.C.L. Leal, and A.C. Rouiller, "Mixed integer programming helping requirements allocation for the NRP in SCRUM teams," in *Proceedings of the XVII Brazilian Symposium on Software Quality*, SBQS '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 279–286.

[S137] F. Sher, D.N.A. Jawawi, R. Mohammad, M.I. Babar, R. Kazmi et al., "Multi-aspects intelligent requirements prioritization technique for value based software systems," in *Intelligent Technologies and Applications*. Singapore: Springer, 2020, pp. 357–371.

[S138] A.K. Kakar, "How does the value provided by a software product and users' psychological needs interact to impact user loyalty," *Information and Software Technology*, Vol. 97, May 2018, pp. 135–145.

[S139] S. Thew and A. Sutcliffe, "Value-based requirements engineering: Method and experience," *Requirements Engineering*, Vol. 23, No. 4, Nov. 2018, pp. 443–464.
[S140] V.C. Gerogiannis and G. Tzikas, "Using fuzzy linguistic 2-tuples to collectively prioritize software requirements based on stakeholders' evaluations," in *Proceedings of the 21st Pan-Hellenic Conference on Informatics*, PCI '17. New York, NY, USA: Association for Computing Machinery, Sep. 2017, pp. 1–6.
[S141] M. Svahnberg and T. Gorschek, "A model for assessing and re-assessing the value of software reuse," *Journal of Software: Evolution and Process*, Vol. 29, No. 4, 2017.
[S142] M. Sadiq, T. Hassan, and S. Nazneen, "AHP_GORE_PSR: Applying analytic hierarchy process in goal oriented requirements elicitation method for the prioritization of software requirements," *3rd International Conference on Computational Intelligence and Communication Technology (CICT)*, 2017.
[S143] V. Freitas, M. Perkusich, E. Mendes, P. Rodríguez, and M. Oivo, "Value-based decision-making using a Web-based tool: A multiple case study," in *24th Asia-Pacific Software Engineering Conference (APSEC)*, Dec. 2017, pp. 279–288.

Appendix B. Value definitions

Value definition(s)	Study ID	Reference(s) used
Value is defined at broader level as: "relative worth, utility or criticality'' or "something intrinsically desirable.''	[S9]	[1]
Value is defined as "a belief that a specific mode of conduct or end-state is personally or socially preferable to its opposite. Values serve as criteria for judgment, preferences, choices, and decisions as they underlie the person's knowledge, beliefs, and attitudes". – p. 74	[S15]	[28]
"Value concerns important benefits of stakeholders, e.g., tangible or intangible, economic or social, monetary or utilitarian."	[S20]	–
"Value can be: profits (generated from products), strategic positioning in market share, utility, relative worth, reputation, customer loyalty, innovation technology, cost reduction, quality of life, or improved productivity." – p. 450	[S25]	[S126]
"Value includes product, process and resource attributes. Value attributes include: profits (generated from products), strategic positioning in market share, utility, relative worth, reputation, customer loyalty, innovation technology, cost reduction, quality of life, improved productivity." – p. 287	[S26]	–

Value is defined based on Theory of Value, i.e., value refers to economic worth of goods and services, and that the value of entities can be seen in different perspectives, e.g., from intrinsic, subjective, or objective angle. Stakeholders have their value propositions and the value can be viewed from different perspectives in different dimensions (e.g., economics, organizational, technical, personal.)	[S38]	[S48], [S86]
"Value is defined as the net worth, or the difference between the benefits and the costs of the asset, all adjusted appropriately for risk, at a given point in time. When the costs are disregarded, implicit, or have been incurred before the point at which an asset is evaluated then value may refer to future benefits or the remaining worth of the asset at that point." – p. 42	[S47]	–
"Value" as "relative worth, utility, or importance." – p. 7	[S49]	–
"The economic concept of value is most commonly defined as the amount of money that a unit of goods or services is traded for. Utility, on the other hand, is all the good and desirable that is created by consuming a product or a service. Hence **the concept of value in VBSE is closer to economic utility than economic value**. To avoid confusion with the terminology, we use the term 'value' for value in VBSE context, and 'economic value' when discussing the economic concept." "In this paper we omit the philosophical definition of value and assume that value exists, and we can use any definition that suits our needs. Hence, we rather ambiguously define **value is the degree of desirability**."	[S77]	[49]
"Value is a measure – usually in currency, effort or exchange, or on a comparative scale – of software (set of programs, procedures, algorithms and its documentation) goods or services that will meet the user's needs, desires, and expectations. All goods or services are being influenced by the quality attributes of the software product." – p. 76	[S95]	[50] [S30]
Value is defined from 3 perspectives: product value (market value of the product, i.e., exchange value), **customer's perceived value** ("benefit derived from the product and is a measure of how much a customer is willing to pay for it, aka use value"), and **relationship value** (between company and customer).	[S101]	Economic theory (no reference)
Proposed definition of value: "The degree of fulfillment of stakeholder's requirements in order of their priority while maintaining the agreed upon commitments and constraints of quality."	[S110]	N/A

"Value depends on the relationship between customer needs and the benefits of products that satisfy those needs." "The value of a product for a customer is expressed in terms of benefit and cost, whereas to a software company it is expressed in terms of the profit (return) from the product sold."	[S126]	[1]
"Values are personal attitudes or long-term beliefs which may influence stakeholder functional and non-functional requirements."; "values may also be interpreted as a set of issues which are frequently referred to as problematic in the RE process, such as politics, culture, sensitivities about the consequences of automation and conflicts between stakeholders."	[S139]	[51]
"A wider view of value that exceeds its economic focus by including aspects such as 'relative worth, utility, or importance', and also presented the concept of key stakeholder to refer to all stakeholders who need to participate in the system definition and development processes."	[S143]	[1]

Appendix C. Detailed list of research topics

SWEBOK knowledge areas	Topics of investigation – Study ID
Software Requirements (34 studies)	Requirements analysis on market-driven software development – [S7]
	Requirements prioritization – [S8], [S13], [S14], [S19], [S24], [S28], [S30], [S35], [S55], [S60], [S65], [S72], [S116], [S119], [S137], [S140], [S142]
	Requirements elicitation – [S15], [S18], [S61], [S62]
	Tool support selection – [S21]
	Requirements tracing – [S86], [S92], [S130]
	Requirements selection decision-making criteria – [S97], [S107], [S126]
	Requirements negotiation – [S99], [S109]
	Value aspects of RE – [S101], [S104]
	Stakeholder identification and quantification – [S103]
SE Management (30 studies)	Software product innovation – [S3]
	Customer value – [S9], [S77]
	Planning and control – [S20], [S32], [S96], [S124]
	Stakeholders' value – [S43], [S108]

SE Management (30 studies)	Decision value/analysis – [S54], [S94], [S64], [S112], [S117], [S143]
	Software project management – [S4], [S25], [S38], [S68], [S136]
	Risks management – [S29], [S102], [S135]
	Product value estimation – [S79], [S95], [S110], [S34]
	Feature usage – [S84], [S88]
	Software value analysis – [S90]
SE Economics (17 studies)	VBSE Agenda – [S1], [S49]
	Product lines – [S36]
	Software initiatives valuation – [S47]
	VBSE Elements – [S50]
	Decision making support system – [S51]
	Value-based monitoring and control (Earned value management system) – [S70], [S105], [S121]
	Return-on-investment model – [S23], [S111]
	Software assurance investment – [S22], [S75]
	Value creation in agile projects – [S115], [S127]
	Multi-dimensional cost analysis – [S134]
	Maintenance cost estimation – [S106]
SE Process (14 studies)	Software process tailoring – [S5], [S87], [S120]
	Software development process – [S11], [S118]
	Software process modeling – [S17]
	Software process improvement – [S27], [S44]
	Process framework – [S31], [S42], [S76]
	Software dependability – [S41]
	Customer value analysis – [S82], [S138]
Software Quality (13 studies)	Stakeholders' alignment – [S10], [S91]
	Software quality assurance – [S16], [S113], [S114], [S123], [S125]
	Software quality processes – [S53], [S57], [S74]
	Verification and validation – [S71], [S73], [S100]
Software Design (11 studies)	Software architecture evaluation – [S12], [S132], [S133]
	Design decision – [S58], [S66], [S80], [S85]
	Components negotiation and reuse – [S63], [S78], [S129]

	System architecture investment – [S89]
Software Testing (08 studies)	Coverage measurement – [S6]
	Usability evaluation – [S45]
	Test management – [S46]
	Testing method – [S56], [S93]
	Testing decision prioritization using machine learning – [S59]
	Testing planning and controlling – [S83]
	Testing prioritization strategy – [S98]
Engineering Foundations (05 studies)	Business value analysis – [S39], [S33]
	VBSE Theory – [S48], [S69]
	Software productivity metric – [S81]
SE Models and Methods (05 studies)	Machine learning in VBSE – [S26]
	Inconsistent stakeholder value proposition – [S37]
	Decision support system development – [S52], [S139]
	Big data value engineering – [S122]
SE Professional Practice (03 studies)	Educational games – [S67]
	Group process – [S2]
	Stakeholder values and conflicts – [S128]
Software Maintenance (02 studies)	Software maintenance impact analysis – [S40]
	Migration of legacy applications and data structures – [S131]
Software Construction (01 studies)	Reuse value assessment – [S141]

Appendix D. Bubble plots and chart

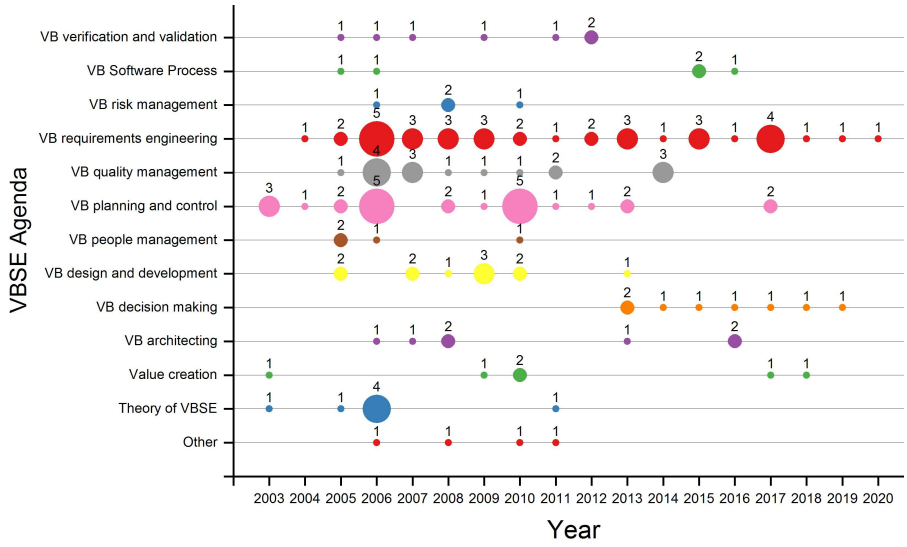

Figure D1. Bubble plot for VBSE principles and practices papers per year

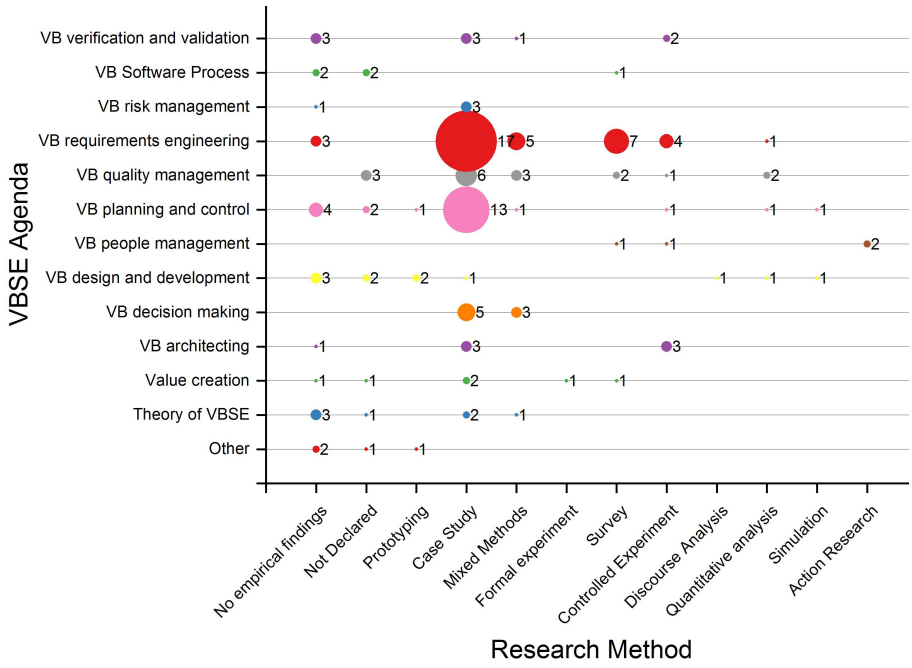

Figure D2. Bubble plot for VBSE agenda vs research method

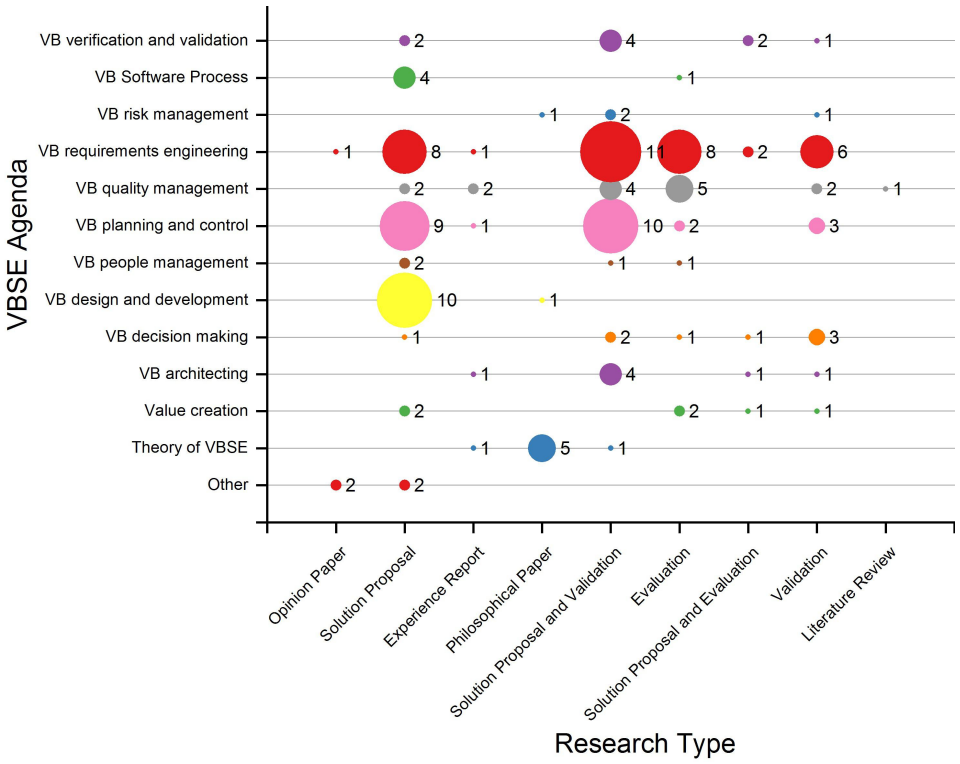

Figure D3. Bubble plot of VBSE agenda by research type

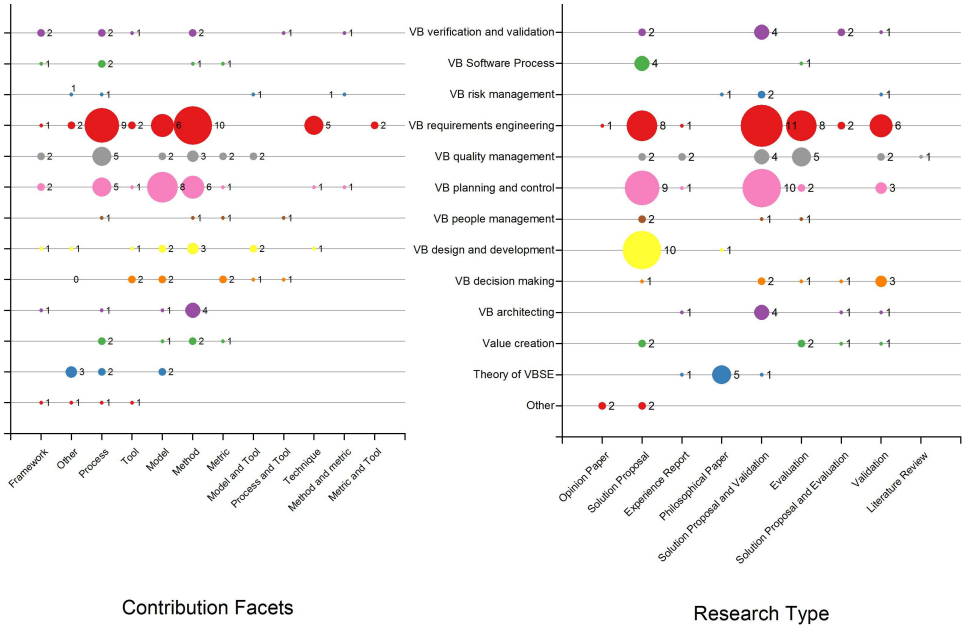

Figure D4. Bubble plot of VBSE agenda by contribution facets and research type

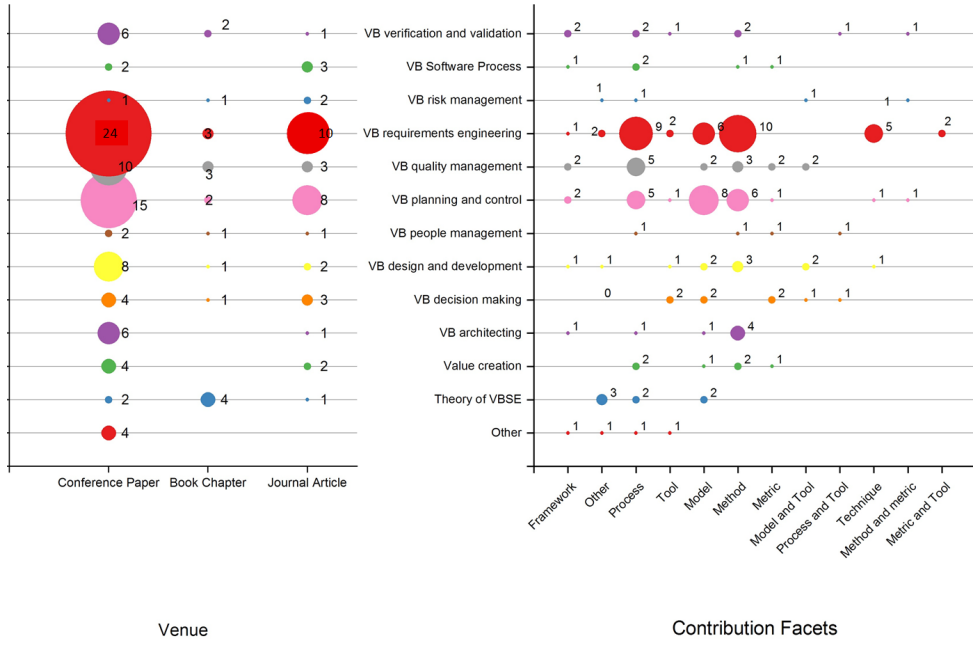

Figure D5. Bubble plot of VBSE areas by publication venue and contribution facets

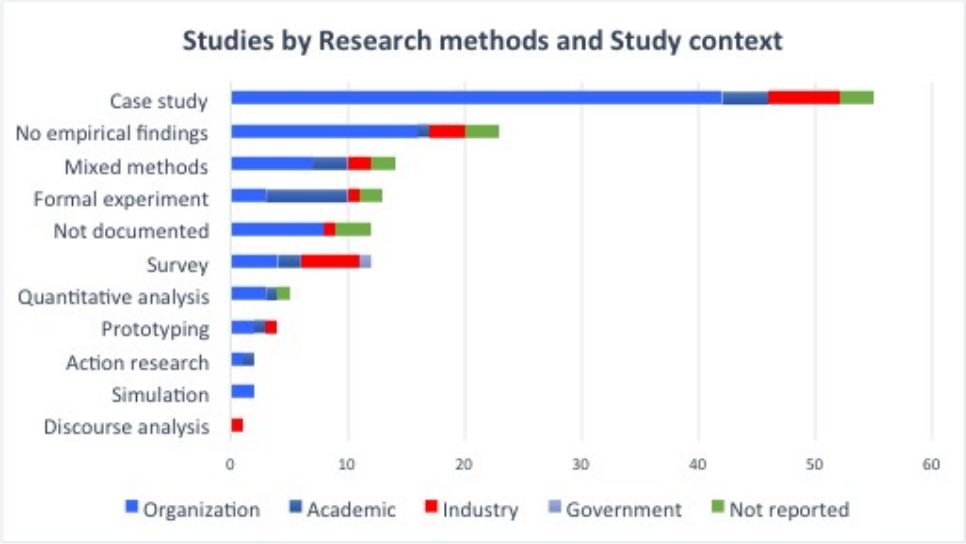

Figure D6. Bar chart on research method by study context

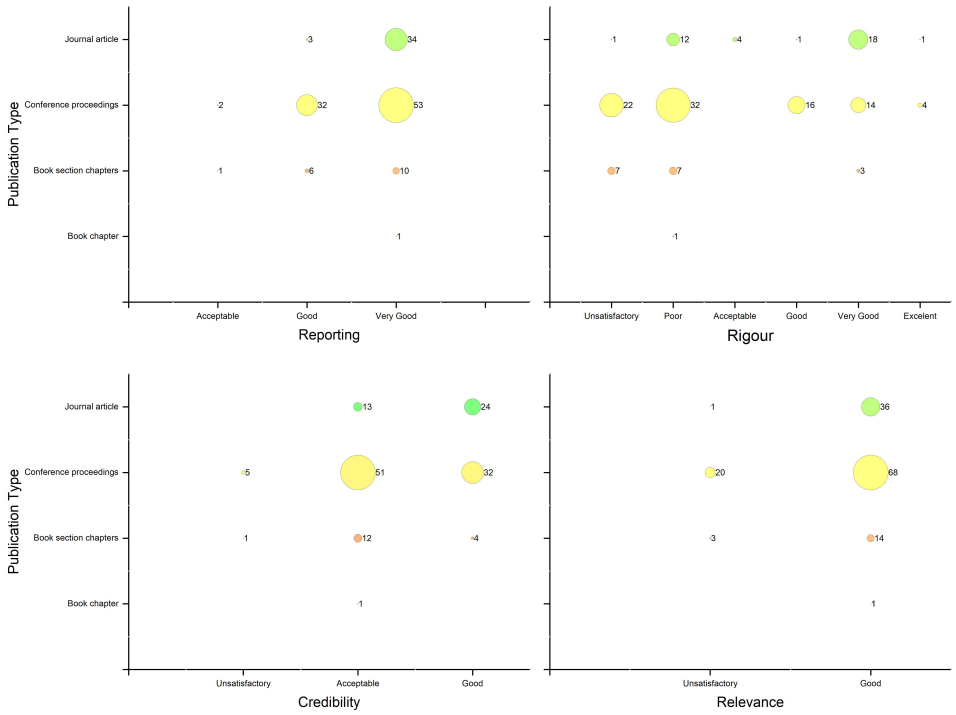

Figure D7. Bubble plot of studies' quality by publication type

e-Informatica Software Engineering Journal, Volume 17, Issue 1, 2023, pages: 230107, DOI: 10.37190/e-Inf230107

Empirical Study of the Evolution of Python Questions on Stack Overflow

Gopika Syam*, Sangeeta Lal*, Tao Chen**
*School of Computer Science and Mathematics, Keele University, UK
**Department of Computer Science, Loughborough University, UK
gopikasyam01@gmail.com, s.sangeeta@keele.ac.uk, t.t.chen@lboro.ac.uk

Abstract

Background: Python is a popular and easy-to-use programming language. It is constantly expanding, with new features and libraries being introduced daily for a broad range of applications. This dynamic expansion needs a robust support structure for developers to effectively utilise the language.
Aim: In this study we conduct an in-depth analysis focusing on several research topics to understand the theme of Python questions and identify the challenges that developers encounter, using the questions posted on Stack Overflow.
Method: We perform a quantitative and qualitative analysis of Python questions in Stack Overflow. Topic Modelling is also used to determine the most popular and difficult topics among developers.
Results: The findings of this study revealed a recent surge in questions about scientific computing libraries pandas and TensorFlow. Also, we observed that the discussion of Data Structures and Formats is more popular in the Python community, whereas areas such as Installation, Deployment, and IDE are still challenging.
Conclusion: This study can direct the research and development community to put more emphasis on tackling the actual issues that Python programmers are facing.

Keywords: Python programming, Software Development, Stack Overflow, Topic Modelling

1. Introduction

Python is a widely used, general-purpose programming language among developers. The language is constantly ranked as one of the most popular programming languages[1,2]. Additionally, according to a recent Stack Overflow survey, the language surpassed Java, C, C++, and SQL to rank third most commonly used programming language[3]. It is a high-level, open-source, user-friendly language developed with a focus on improving code readability and development speed [1]. Python Programming language has undergone tremendous changes over years, and new language features are constantly being added to it [2].

[1]https://pypl.github.io/PYPL.html
[2]https://www.tiobe.com/tiobe-index/
[3]https://insights.stackoverflow.com/survey/2021#overview

Submitted: 21 Jan. 2023; Revised: 7 Jun. 2023; Accepted: 27 Jun. 2023; Available online: 4 Jul. 2023

Python has a large collection of libraries, frameworks, and integration support, making it incredibly useful for programmers. The language is constantly evolving, with language extensions, performance improvements, and core library updates. Because of these libraries and functionalities, it can be used in a wide variety of areas, making it popular among developers. However, developers are generally expected to be experts in these libraries and to maintain their skill set up to date, and keeping up with the pace of language development is often difficult for them. As a result, they frequently rely on programming question-and-answer sites like Stack Overflow[4] to ask for help on the challenges they encounter.

The development community commonly uses Stack Overflow, one of the major code--focused websites that is a part of the stack exchange network[5], to post challenges and exchange ideas. Stack Exchange is a network of question-and-answer websites that features discussions on various subjects in multiple fields. Some of the most well-known Stack Exchange websites include Stack Overflow[6], Super User[7], Server Fault[8], Ask Ubuntu[9], Data Base Administrator[10], Android User[11], Software Engineering[12] and others. Each of these websites features conversations on various topics and enables users to post questions and receive responses. Voting on questions and answers allows users to gain reputation points, which in turn grants them more privileges. Users with higher privileges can provide comments on questions and serve as moderators for certain sections of the website.

The most actively viewed sites in this network are Stack Overflow, Mathematics, Unix and Linux and Ask Ubuntu as of February 2023[13]. Stack Overflow has over 100 million monthly visits and has received around 21 million questions and 31 million answers to date[14], and it serves as a platform where developers can get expert advice on a variety of issues related to software development. With the help of this study, we are given a broad and realistic picture of developers' discussions about Python, helping us to learn real-world knowledge about the volume of these discussions and the topics that developers find both popular and difficult.

1.1. Motivation

Python is always expanding and evolving, with more developers adopting the language for diverse applications and new libraries and frameworks being released for a wide range of uses. Like other languages, the frameworks in Python are evolving constantly, mainly due to feature enhancements, performance improvements, and bug fixes. In spite of Python's popularity, researchers have not focused much attention on analyzing the trends and technologies of this language on Q&A websites like Stack Overflow [3]. Analyzing such Q&A platforms can provide insights that can help in making better tools and technologies for Python developers.

[4]https://stackoverflow.com/
[5]https://stackexchange.com/
[6]https://stackoverflow.com/
[7]https://superuser.com/
[8]https://serverfault.com/
[9]https://askubuntu.com/
[10]https://dba.stackexchange.com/
[11]https://android.stackexchange.com/
[12]https://softwareengineering.stackexchange.com/
[13]https://stackexchange.com/sites?view=list#traffic
[14]https://stackoverflow.co/

Python released a new version of the language in 2008 that is not backwards compatible, causing a transitional phase for Python developers. One major issue is that programs written in older versions of the language cannot be interpreted in the new version of the language without modifications. This is a challenge for developers because they must update or rewrite their applications to work with the latest language version. The study by Malloy and Power [4] discussed in detail the impact of this transition from Python 2 to Python 3 on the applications written in Python.

Table 1: Example of Python questions from Stack Overflow website

S. No.	QId	Title
1.	30 667 525	ImportError: No module named sklearn.cross_validation
2.	38 987	How do I merge two dictionaries in a single expression in Python?

In addition, we can observe that developers frequently struggle with package name changes, which result in import and module not found errors, which is a roadblock in the early phases of development. Consider the Question Id (QId): 34 844 352[15] in Table 1, with 440 000 views where the user is experiencing "Import Error" with "sklearn.cross_validation" module. The issue in this question is because of the renaming and deprecation of the "cross_validation" sub-module to "model_selection" and it took approximately seven months to get an accepted answer. A lot of similar questions can be seen in Stack Overflow related to the deprecation of packages. Also, there are multiple Python questions in Stack Overflow, that took plenty of time to get a response. For example, the Question with Id: 38 987[16] in Table 1 has 2.9 million views and took almost 6 years to get an accepted answer. It can be challenging for programmers to create and maintain Python applications due to this constant evolution and dynamic characteristics, which can also affect the language's effectiveness, safety, and development time.

It is crucial to address these concerns and learn more about the actual difficulties faced by the Python development community in the real world. Some recent studies highlight the importance of analyzing issues faced by Python developers. A study by Zhang [5], investigates in detail the typical patterns and examples of issues faced in the evolution of Python APIs. This shows how Python framework evolution may result in compatibility issues in client applications. Seeing the importance of Python in the software development community, Widyasari et al. [6] recently proposed a Python bugs dataset.

The work presented in this paper is complementary to the above studies. In this work, we analyze Python questions from the Stack Overflow website to understand more about the evolution of the Python language and developers' challenges. We believe that the broad and diversified user base of the platform offers the ideal setting for investigating how various programmers use Python in their projects and what are some of the popular and challenging topics of discussion. This study will aid in establishing the extent to which current research in the Python language is deficient and understand the major causes for this by examining real-world difficulties thereby paving the way for future research in this field. We performed a multi-dimensional study of the Python questions on Stack Overflow. We analyzed the evolution of questions to understand the growth of the Python questions and answers over the period of time. We analyzed popular tags associated with

[15] https://stackoverflow.com/questions/30667525/importerror-no-module-named-sklearn-cross-validation/34844352#34844352

[16] https://stackoverflow.com/questions/38987/how-do-i-merge-two-dictionaries-in-a-single-expression

the Python questions to understand how interest in technologies changed over the period of time. Next, we identify popular topics (as well as their popularity and difficulty analysis) in Python questions to understand the common issues faced by Python developers. We believe that such analysis can be beneficial in revealing the most interesting and difficult areas of Python software development.

1.2. Goal and Research Questions

The objective of this study is to *conduct an in-depth analysis of the Python post in Stack Overflow to understand the evolution of Python and gain a better understanding of the challenges faced by developers*. We believe that the findings of this study can be utilised by practitioners, researchers, and educators in understanding the current state of Python in terms of challenges and trends and the extent to which the traditional viewpoint is different from real-world applications and issues. Furthermore, the study can also aid in identifying the areas where additional language resources and tools are required to support developers.

We begin by assessing the volume of discussion about Python and its distribution on the Stack Overflow platform to determine how popular the Python language is within the community. Then, user contributions and response times are investigated further to determine the availability of experts to address the challenges encountered by developers. We also assess the tags attached to the Python question and also identify the main topics of discussion. Finally, we categorise the topics based on several indicators to identify the popular and difficult topics. In conclusion, the following research questions are the focus of our study.

RQ 1: How have Python posts evolved throughout the years? The first research question (RQ 1) examines how the Python questions or posts have changed over time. This RQ is divided into four smaller RQs, each of which examines a distinct feature of Python posts that have been published on the Stack Overflow platform over time. The first sub-RQ (RQ 1.1) aims to gain insights into the growth of Python posts over the years by exploring the volume of questions and answers in the Stack Overflow platform. The second sub-RQ (RQ 1.2) tries to understand the amount of discussion in the platform and the satisfaction and difficulties of developers by exploring the questions with accepted answers, non-accepted answers, comment-only questions, and questions with no answers or comments The third sub-RQ (RQ 1.3) investigate how long it takes for Python questions to receive a response or an accepted response to comprehend the user's satisfaction and the availability of experts on the platform. The fourth sub-RQ (RQ 1.4) tries to understand the participation of the development community by determining the user's posting questions, answers, and accepted answers.

We observed an increase in the number of questions across the years, except in 2021, and a decrease in the percentage of questions receiving accepted answers. A huge satisfaction among the Python development community was also observed with most questions getting answered within the first hour of asking the questions.

RQ 2: How did the Python tags evolve over the years? The questions in Stack Overflow have tags assigned to them by the user. This RQ analyses these tags to determine the tags with the greatest number of questions. The findings showed that the popularity of data analytics and artificial intelligence is rising indicated by the increase in the number of questions in the "pandas" tag. It was also observed that Python developers still struggle with general programming questions as most of the question tags belonged to this category.

RQ 3: What are some of the main topics of Python discussed by Python developers? Every Python post on Stack Overflow reflects the difficulties programmers encounter when learning and creating Python applications. We extracted semantic concepts from these posts using topic modelling to better comprehend this. There were nine topics, including the general programming topic (e.g., algorithm implementation, object-oriented principles, general Python concepts, etc.), Scientific computing topic (data format handling, model implementations, big data libraries, etc.), Web development topic (mostly linked to Django and flask libraries), and others. We also explored some of the major discussion points and questions in each of these topics to understand more about the problems faced by developers in each of these topic categories.
RQ 4: What are the most popular and challenging topics discussed by developers in Stack Overflow? In this RQ we evaluated the Python topics identified in the previous RQ to determine the popular topic and the topic that is difficult to answer on Stack Overflow. Many metrics are used to categorise the topics into these categories as per prior research. The interesting observation in this RQ was that the topic of Installation, Deployment, and IDE was popular as well as challenging for developers.

In the context of this paper, the terms "Python questions" and "Python posts" are used interchangeably, indicating that they refer to the same concept. The remainder of this paper is organized as follows: Section 2 describes the study methodology used in our study and offers an in-depth discussion of the various steps involved. Section 3 provides the results and analysis for each of the research questions in this study, as well as information about the motivation, approach, and results for each. The major discussion points and insights this study offers to the research, educational, and development communities are presented in Section 4. Section 5 discusses several risks to validity, and, Section 6 discusses related works that use the Stack Overflow dataset to evaluate various aspects of programming languages, concerns encountered in Python software development, and software development challenges. Section 7 finishes the study with the conclusion and a brief discussion of future research.

2. Study methodology

Both quantitative and qualitative data analysis is used in this study to analyze different facets of Python programming [7]. This is also known as the mixed-method approach [7]. Using both of them offers valuable insights into the relationship between qualitative and quantitative data [7]. To be more explicit, the process employs several well-known statistical approaches, such as unsupervised machine learning and natural language processing to the dataset to post trends and patterns in Python posts. To support and correlate our quantitative findings, a manual analysis of a statistically significant sample of Python posts is conducted to learn more about the difficulties faced by Python developers. This practice is used by other researchers in the software engineering domain [8].

Figure 1 shows the methodology of our study. It consists of four main activities: (1) Obtaining the current Stack Overflow dataset, (2) Extracting the Python tags required for Python posts extraction, (3) Extracting Python posts or queries relating to the Python programming language (i.e., questions, answers, metadata), and (4) analyzing the metadata and textual content in the Python posts. In summary, we use the Stack Overflow dataset dump provided by *The Internet Archive*[17]. From this dataset, we extracted all the Python

[17]https://archive.org/download/stackexchange

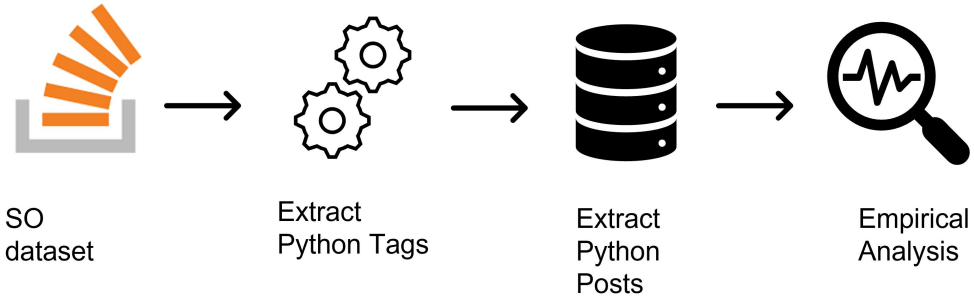

Figure 1: Overview of study methodology

posts (using tags of the post) and analyze these posts using both automated and manual processes to answer the research questions. The detailed steps related to each activity are mentioned in the following subsections.

2.1. Stack Overflow dataset extraction

The data for this study was gathered from Stack Overflow[18]. Stack Overflow is a website with over 14 million registered users that allows them to post questions and answers about many aspects of computer programming. It also allows users to edit their questions and answers, upvote or downvote questions and mark accepted answers. Users are also given badges and reputations based on the number of upvotes they receive and their meaningful contributions to the platform, which allows them to gain access to extra features like commenting and editing other people's posts[19]. The Stack Overflow data used for this study is obtained from *the internet archive*, which hosts the data of all Stack Exchange forums[20]. The platform has 8 files related to Stack Overflow: *badges.xml, comments.xml, postHistory.xml, postLinks.xml, post.xml, tags.xml, user.xml*, and *votes.xml*. The data was downloaded from this forum on June 2022. Figure 2 shows a snapshot of a post from Stack Overflow website[21]. Following is a brief description of the Stack Overflow dataset components:

Posts: There are two types of posts on Stack Overflow: (1) Questions and (2) Answers. The question post consists of a problem or challenge faced by the developer. A question has a title and body part. The title is plain text that gives the summary of the problem faced by the developer whereas the body part is a detailed description of the problem faced by the developer. The developer can put sample code, stack traces, etc., to provide more details about the problem. The Stack Overflow community can provide a response to the question. The response is called an "answer" post. There can be multiple answer posts associated with a single question post. The answer post can be of two types: Accepted answer or non-accepted answer. The accepted answer shows that the developer who asked the question considers this answer posts a solution to the question asked. Only the developer who asked the question can mark an answer as the accepted answer.

Comments: User can also provide a response to a post by posting comments on it. The users post comments on a post when they need clarification from the author of the post,

[18]https://stackoverflow.com/

[19]https://en.wikipedia.org/wiki/Stack_Overflow

[20]https://archive.org/download/stackexchange

[21]https://stackoverflow.com/questions/240178/list-of-lists-changes-reflected-across-sublists-unexpectedly

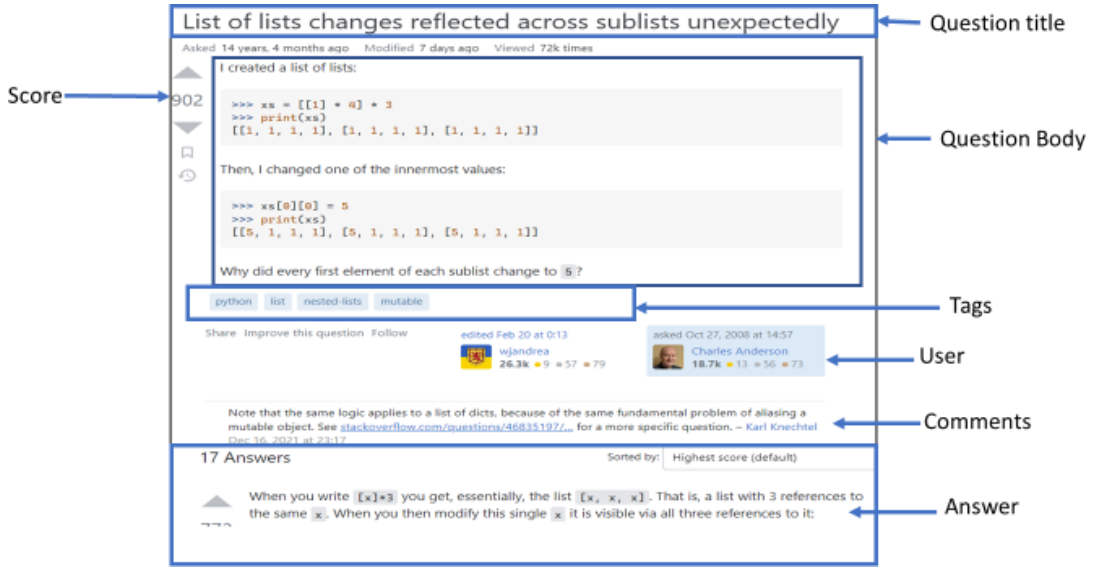

Figure 2: A snapshot of a post from Stack Overflow

they want to leave constructive criticism that is helpful for the author to improve the post, or when they want to leave some minor information about the post.

Tags: A developer is required to associate one to five tags with the question (while creating any new question). These tags are useful in categorizing the questions. It helps the developer community on Stack Overflow to search questions that belong to a specific category. In Figure 2, the questions have four tags – "Python", "list", "nested-list", and "mutable"

Score: The Stack Overflow community can upvote or downvote a post based on their analysis of how useful a particular post is for the community. The *score* metric shows the total number of upvotes minus the total number of downvotes. In Figure 2 the score of the question is 902.

View Count: This metric shows how many times a question post is viewed. It is associated with only question posts. In Figure 2 the question is viewed 72k times.

Favorite count: This metric shows how many times the developers on the Stack Overflow community marked a post as their favourite. The metric is only associated with question posts.

User details: The post also has information about the user who asked the questions, answered the question, commented on the post and edited the post.

2.2. Python tags and posts extraction

The dataset contains information about all the posts on the Stack Overflow website. For this study, we focused our attention on discussions related to Python. Hence, we need to extract all the relevant posts for this study from the obtained *internet archive* dataset. Python is a popular programming language, therefore manually collecting Python questions from this vast dataset is challenging. So, to identify all the Python-related questions, we followed the steps mentioned below:

First, we created a list of all Python-related tags. We used data exchange explorer[22] and extracted the top 5000 questions with the highest score with the tag <python>. We used

[22]https://data.stackexchange.com/stackoverflow/query/edit/1600435#resultSets

the "score" account as a criterion because it is used by other researchers [8]. Additionally, the score count provides more authentic criteria for the usefulness of a question than the view count. Because when a user search for a question, they may view many questions before finding a question that resolves their problem. This increases the view count of many questions even when it does not provide useful answers. However, the score shows the number of upvotes that a question received. A user/moderator will only upvote a question they think is useful.

All the extracted posts were manually analyzed by the first author of the paper. The first author then created an initial list of Python tags by analyzing the various tags associated with these questions. This involved researching the tags and their relation to Python language. For example, the tag <eventlet> was further researched and its online documentation[23] helped to identify the relation of this tag to Python. During this process, There were some tags for which the first author was sure that they are related to Python. Such tags are put into "**Python-related**" tag list. The tags for which the first author had any confusion were put into another list called the "**doubtful tag**" list. The "**Python-related**" tag list was reviewed by the second author to identify whether these tags are Python related or not. During this review, all the disagreements between the first author and the second author are resolved through discussion. However, if the agreement is not reached for any tag, such tags are moved to the "doubtful-tag" list. Using this process, we identified 169 tags.

Second, we performed a second review of all the tags present in the "doubtful-tag" list. We noticed that most of the tags in this are either too generic or not popular. For each tag in the list, we extracted the top 30 highest-score question posts. The first author reviewed all 30 questions for each tag and either moved the tag to the "Python-related" tag list (if at least 90% of the 30 questions analyzed for a tag were related to Python) or kept it in the "doubtful-tag" list (in case of confusion). Both of these lists were again verified by the second author. We kept a tag in the "Python-related" tag only if both the first and the second author agreed on a tag to be Python-related. If there is a disagreement with respect to any tag it is moved to "doubtful-tag-list". For example, we analyze 30 question posts containing the tag <tuples>. We noticed that the out of these 30 question posts 21 questions were related to Python, 3 were related to C#, 4 were related to C++, and 1 was related to Java and typescript. So, this tag was kept in the "doubtful-tag" list. On the other hand, tags like <psycopg2> and <pep8> were added to the "Python-related" tag list after observing that all 30 questions manually checked were related to Python. Using this process, we added 5 more tags to the "Python-related" tag list and collected a total of **174** tags. These tags had 100% agreement between the first author and the second author because all the tags in which there was a disagreement were removed from this and moved to the "doubtful-tag-list" (see Table 2).

Third, previous studies on the Stack Overflow dataset mentioned the use of the title of questions (in addition to tags) for extracting the desired questions [8]. Hence, we explored titles of questions to extract more Python questions. For this, we extracted the 50 questions with the highest scores that have the word "Python" in the title and did not include any of the tags from our "Python-related" tag list. The content of these 50 questions was analyzed manually by the first author to determine their relevance to Python. Out of the 50 questions, 20 were related to Python while the remaining 30 were related to other programming languages. These questions' posts were further reviewed by the second

[23]https://eventlet.net/

Table 2: Final tags included in our Python-related tag list

List of Tags
Python, python-module, python-class, python-datamodel, pandas, python-3.x, python-packaging, matplotlib, python-internals, pip, python-import, pyenv, python-venv, pypi, mysql-python, zen-of-python, python-c-api, python-multithreading, python-2.x, python-unicode, setup.py, python-os, pyc, Django, django-models, psycopg2, python-wheel, ipython, python-typing, python-requests, cherrypie, cpython, django-queryset, pandas-loc, python-multiprocessing, python-2.5, python-2.7, python-zipfile, python-datetime, pandas-groupby, ironpython, pytest, python-3.6, flask, django-views, numpy-ndarray, numpy, python-imaging-library, python-2.6, beautifulsoup, timedelta, urllib2, scipy, seaborn, pythonpath, django-orm, pyyaml, python-nonlocal, python-unittest, pylint, pyflakes, pychecker, python-3.3, django-admin, pywin32, gunicorn, pytorch, mypy, pickle, django-shell, shutil, fnmatch, django-templates, wxpython, numpy-einsum, imaplib, flask-sqlalchemy, pygtk, pyspark, python-mock, flake8, django-media, django-authentication, python-asyncio, python-3.5, pycurl, python-3.4, django-signals, itertools, scikit-learn, scrapy, python-logging, eventlet, django-2.0, tkinter, openpyxl, pprint, django-testing, pyperclip, smtplib, popen, pypdf2, pypdf, configparser, django-validation, django-urls, python-sphinx, pycrypto, python-control, pymongo, django-manage.py, manage.py, pandas-explode, nonetype, django-migrations, paramiko, python-idle, python-dataclasses, gevent, django-rest-framework, nltk, pytz, difflib, distutils, py2exe, python-poetry, django-custom-manager, python-attrs, google-api-python-client, python-closures, pyarrow, django-aggregation, django-staticfiles, pyinstaller, django-class-based-views, python-2to3, mod-python, pydev, django-modeltranslation, imghdr, pyqt, magicmock, python-docx, pathlib, numpy-slicing, Theano, python-nose, django-q, django-celery, python-rq, fillna, pyqt4, django-1.10, python-click, h5py, pytables, pygame, jython, jpype, pycharm, jupyter-notebook, anaconda, jupyter, pypy, tensorflow, keras, conda, opencv, pep8, argparse, f-string, timeit, xlrd, mplcursors

author. We noticed that most of the questions were not related to Python but rather were demanding Python functionality in other programming languages. We opted not to include these questions in our final extracted questions because of the significant number of false positives. A sample from these questions is shown in Table 3.

Table 3: Illustrates the questions with a Python tag in the title, but are related to other languages

Programming language	Title	Question Id
R	Does R have an assert statement as in Python?	2 233 584
Java	Java plotting library like Python's matplotlib	958 806
C++	C++ algorithm like Python's "groupby"	12 335 860
Node.js	Node equivalent of "python -m SimpleHTTPServer"?	22 513 544
Javascript	Is there a javascript equivalent of Python's ___getattr___ method?	1 529 496

Fourth, we looked at 30 questions with the highest score that had the word "Python" in their body but neither consist of any of the identified tags from our "Python-related" tag list nor the word "Python" in their title. Each of these questions was manually verified to confirm their relation to Python. Analysis revealed that all 30 questions containing the word "Python" in the body were not actually related to Python programming language. Nine of the questions were general inquiries about programming concepts, coding frameworks, keyboards, and other topics, while 21 questions pertained to other programming languages.

For instance, a list comprehension method in ruby, similar to the one in Python is requested in Question Id: 52 624[24]. Due to the enormous number of false positives, we decided to exclude the question with "Python" in the body alone from our analysis. A similar approach is followed by other researchers as well [8].

Table 4: Summary of collected data

Item	Value
Total number of Python questions	2 449 567
Question with Accepted Answers	1 231 993
Percentage of questions with accepted answers	50.29%
Timestamp of the first question	2008-08-02 03:35:56
Timestamp of the last question	2022-06-05 06:38:44
Total tags related to Python	174

Fifth, we extract all the question posts consisting of at least one tag from our "Python-related" tag list which has 174 tags. Using this process we extracted 2 449 567 questions. We observed that The number of questions extracted by our approach is comparable to the number of questions extracted in other studies [3]. The details about our extracted dataset are mentioned in Table 4.

2.3. Results analysis

We perform both quantitative and qualitative studies of Python posts. For quantitative analysis, we use tools such as SQL queries and code scripts to extract the information from the dataset. Whereas, for the qualitative study we perform manual analysis of the dataset. Our quantitative strategy includes running database queries and performing topic modelling analysis with an unsupervised machine learning algorithm. The qualitative approach, on the other hand, involves authors manually analysing statistically significant sample sets of data.

Table 5 shows the details about each research question and the methodology (quantitative or qualitative) we followed to answer it. For instance, in RQ 3, we first use a quantitative technique to identify the topics using LDA analysis, and then we use a qualitative strategy to analyse a sample of questions within each topic given by LDA to correctly label it. In Section 3 we provide more details about the research questions and the corresponding results obtained.

Table 5: An outline of RQs and the respective research approach that we used to answer each RQ

RQ Categories	RQ details	Research Approach
Questions Metadata	RQ 1	Quantitative
Tags	RQ 2	Qualitative and quantitative
Topics	RQ 3, RQ 4	Qualitative and quantitative

[24]https://stackoverflow.com/questions/310426/list-comprehension-in-ruby

3. Empirical study design and results

In this section, we provide details of all the research questions. We describe the motivation, approach, and findings of all the research questions.

3.1. RQ 1: How have Python posts evolved throughout the years?

The first research question (RQ 1) examines how the Python questions or posts have changed over time. This RQ consists of four sub-RQs. Each question focuses on a different aspect like the number of Python questions posted annually (RQ 1.1), the trend of successful, unsuccessful, ordinary, and comment-only questions (RQ 1.2), the time to get an accepted answer to Python questions (RQ 1.3), and the number of users who post questions and answers (RQ 1.4).

3.1.1. RQ 1.1: How did the Python post change over the years?

Motivation: This RQ determines the popularity of Python over time by analysing the number of Python questions posted between 2008 and 2022. This will be useful in understanding how the language has evolved over time and to what extent developers struggle with it. Determining the increasing or decreasing trend in the number of questions posted, for example, will help Stack Overflow, Kaggle, and other similar website teams to assign moderators based on this pattern. In this RQ, we use all the extracted Python questions as well as all the questions with accepted answers based on the identified set of tags. A question with an accepted answer typically indicates whether it was helpful. This will be beneficial for assessing Python developers' performance and satisfaction.
Approach: We group all the questions by year and saved them as a key-value pair in the Python dictionary. A line chart is drawn that shows the yearly trend of all the Python questions posted between 2008 and 2022. Another dictionary is used to store the set of questions with accepted answers based on the year. A bar graph is created that shows the yearly trend of total questions posted in the year and the number of questions having accepted answers.
Results: We extracted 2 449 567 Python questions from which 1 231 993 (50.29%) questions had an accepted answer. Table 4 provides the summary of the collected data. Figure 3 shows the yearly breakdown of the number of Python questions and Python questions with accepted answers. In this chart, for each year the blue bar is representing the total number of questions and the green bar is representing the questions with accepted answers. It is interesting to note that the number of questions posted each year is increasing (except for the years 2021 and 2022). This shows and complements the increase in the popularity of Python questions and further motivates this study.

A slight decline can be seen in the number of questions posted in the year 2021. There can be many reasons for that. One reason could be the COVID-19 pandemic. Because of this many sectors faced a decline in the number of jobs posted. Maybe developers working in Python also faced similar issues[25]. Georgiou [9] reported a decline in the number of questions posted during COVID. Another reason can be because of the rise of popularity

[25]https://www.zdnet.com/article/developer-jobs-demand-for-programming-language-python-falls-amid-pandemic/

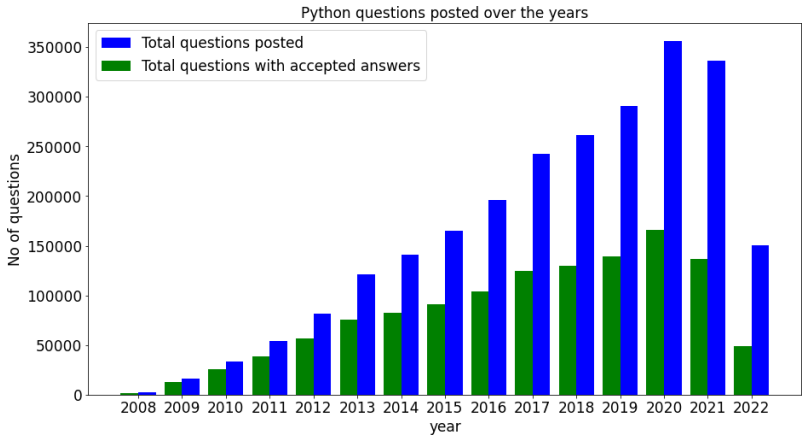

Figure 3: The number of total questions and questions with accepted answers over the years

of Java because of the advancement in Android Apps, SmartTV, Desktop Apps, and many more[26].

In our analysis, we consider data from 2008 to 2022. We notice a decreasing trend percentage of questions with accepted answers, i.e, 2008 has the highest percentage of accepted answers (81.49%) and 2022 has the lowest percentage of accepted answers (32.42%). We believe that this happened because older questions have a higher probability of getting accepted answers in later years.

3.1.2. RQ 1.2: What are the trends of successful, ordinary, comment-only, and unsuccessful questions?

Motivation: In this RQ we analyze successful, ordinary, comments-only, and unsuccessful questions. On Stack Overflow, users can post questions, answers, and comments on the Stack Overflow website. If the user (who asked the question) is satisfied with an answer, she can mark it as accepted. According to the study by Pinto [10] a question with an acceptable answer might be deemed a **successful** question because it indicates the user's satisfaction. An **ordinary** question, on the other hand, has answers but none of them are marked as accepted. An unsuccessful query is one that has no answers.

We frequently observe that the comments section under each question is also an excellent discussion area where users could get some tips or answers. Therefore, to decide the volume of discussion, we can analyse questions with comments and no answers. We called such questions as **comment-only** questions. Questions that have no answers or comments are considered **unsuccessful**. In this RQ, we aim to identify the trend of successful, ordinary, comments-only, and unsuccessful questions to have a better understanding of the satisfaction and difficulties faced by Python developers.

Approach: Out of all the questions, Python-related ones were filtered out based on the identified set of tags. The question with an accepted answer was extracted separately as a key-value pair into a Python dictionary based on their year. All questions with an "answercount" field greater than one and without accepted answers were then aggregated by year as "ordinary" questions. Out of the remaining questions, the ones with "commentcount"

[26]https://gettotext.com/python-java-and-sql-the-most-in-demand-programming-languages-in-2022/#:~:text=Perhaps%20the%20main%20reason%20for%20this%20precipitous%20drop,greatly%20influence%20the%20salary%20you%20can%20ask%20for

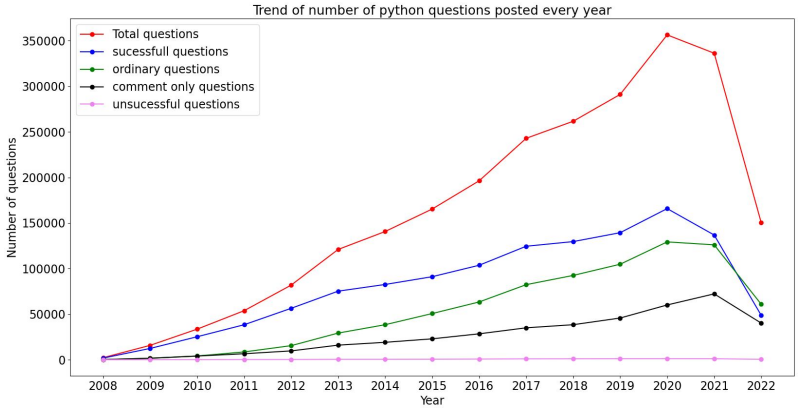

Figure 4: Question distribution over the years

greater than one was then grouped based on year into the dictionary. All the remaining questions were considered unsuccessful and were grouped together.

Results: Figure 4 shows the results obtained for this RQ. We extracted a total of 2 449 567 Python question posts. Out of these, 1 231 993 (50.29%) questions were successful (got an acceptable answer) and 807 735 (32.97%) questions were ordinary questions (i.e., had at least one answer), 400 760 (16.36%) questions are comment-only (i.e., only comments posted for these questions), and the remaining 9079 (0.37%) questions were unsuccessful (i.e., questions neither received any answers nor any comments). Figure 4 also shows that the number of successful questions outnumbers the ordinary, comment-only, and unsuccessful questions over the years (except in 2022). A decline in the year 2022 may be because we don't have complete data from the year 2022 as our experimental dataset was extracted in June 2022. The results show that a large number of questions received an accepted answer or at least one answer. This shows high satisfaction among Python developers. It is also interesting to show that the number of unsuccessful questions is very low over the years, i.e., $\leq$0.57%.

When comparing the results to the earlier study on Java [11] with Stack Overflow data, it can be seen that the total number of questions posted in Java is not always increasing and became more stable during 2013, 2014, followed by a fall in 2015. However, Python questions were consistently increasing during that time period, demonstrating how the language grew in popularity and eventually surpassed Java questions by 2017.

3.1.3. RQ 1.3: How much time it takes to get an accepted answer for Python questions?

Motivation: In this RQ, we analyze how much time it takes Python questions to get the first answer as well as the accepted answer. This is important as insight into the amount of time it takes to get an accepted answer or first response for Python questions can help in providing more evidence about the satisfaction among the Python community and support the results of prior RQs.

Approach To calculate the time for receiving the first answer and accepted answer, we collected a statistically significant sample of questions. We collected 16 529 questions from our experimental dataset with a 99% confidence level and 1% confidence interval [12, 13]. We created this sample using random sampling without replacement. We kept the confidence interval (i.e., margin of error) to 1%, this means that the true results might differ by $\pm$1% from the results obtained. We selected a confidence level equal to 99%. The

confidence level measures the uncertainty regarding how accurately a sample reflects the population being studied within a chosen confidence interval. A confidence level of 99% means that we are 99% certain that the results obtained using the sample match that of the actual population. We used the following formula for calculating the sample size[27]:

$$\text{Unlimited population } (n) = \frac{z^2 \times \hat{p}(1-\hat{p})}{\epsilon^2} \tag{1}$$

$$\text{Finite population } (n') = \frac{n}{1 + \frac{z^2 \times \hat{p}(1-\hat{p})}{\epsilon^2 N}} \tag{2}$$

Here, n – sample size obtained using an infinite population,
n' – sample size for finite population,
z is the z score,
ϵ is the margin of error or confidence interval,
N is the population size,
$\hat{p}$ is the population proportion.

Sample size calculation:

$$\text{Unlimited population } (n) = \frac{2.58^2 \times 0.5(1-0.5)}{0.01^2} = 16\,641 \tag{3}$$

$$\text{Finite population } (n') = \frac{16\,641}{1 + \frac{2.58^2 \times 0.5(1-05)}{0.01^2 \times 2\,449\,567}} = 16\,529 \tag{4}$$

In statistics, If the population is large often the information is inferred by studying a sample of that population. This is a well-known approach for large datasets and has been used by many other researchers in the software engineering domain [12, 13]. Hence, we believe that it is effective in our study as well. For each of these questions in the sample, we calculate the time for receiving the first answer and accepted answer with the help of the "CreationDate" field associated with each post in the Stack Overflow dataset.

Results: From the sample data, we looked at the time at which the answer post was created for each question. The findings showed that most of the responses to the Python questions occur within the first hour. The histogram in Figure 5 shows the distribution of time in hours from when a question is posted until the answer is obtained. This indicates that most of the Python questions receive an answer within a short period implying significant satisfaction within the developer community which further supports the developer satisfaction as identified in the prior RQ.

We also looked at the time at which an accepted answer is obtained. The histogram in Figure 6 shows the distribution of Python questions and reception of an accepted answer. The graph shows that most of the Python questions receive an accepted answer within the first hour. This indicates that there are experts in the Stack Overflow community to provide answers to the challenges faced by the Python software development community.

[27] https://www.calculator.net/sample-size-calculator.html

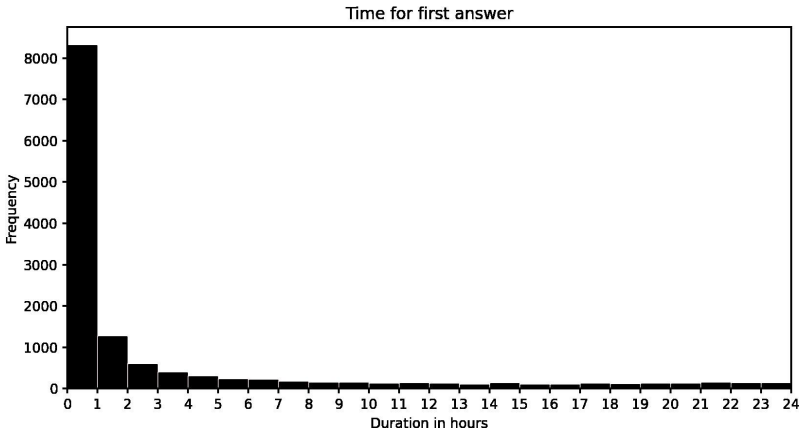

Figure 5: Time distribution of when a user asked a question and received the first answer

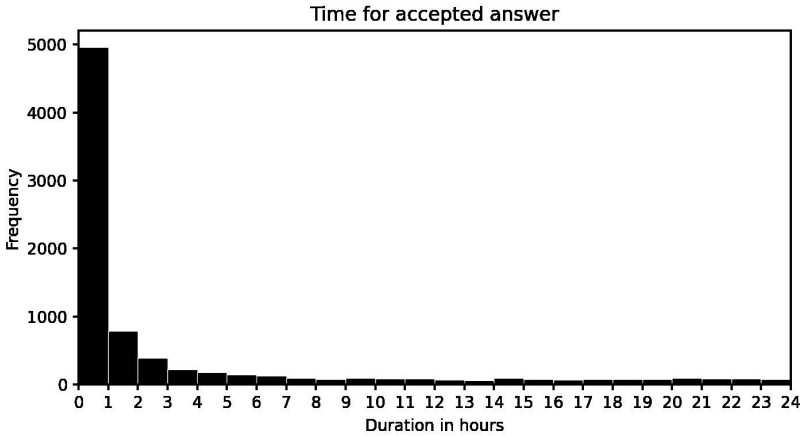

Figure 6: Time distribution of when a user asked a question and received an accepted answer

3.1.4. RQ 1.4: What is the distribution of Python questions and answers among developers?

Motivation: In this RQ, we analyze the Stack Overflow community participation in Python questions and answers. This analysis can provide insights into the size of the Python development community on Stack Overflow as well as whether only a few members of the community are responsible for asking and answering Python questions.
Approach: To determine the users posting questions, answers and accepted answers. we collected a statistically significant sample of questions. We collected 16 529 questions from our experimental dataset with a 99% confidence level and 1% confidence interval [12, 13]. To identify the user posting questions and answers we use the "OwnerUserId" field. We filter out users who posted questions, accepted answers, and non-accepted answers.
Results: The results showed that 15 361 unique users posted Python questions and 16 381 unique users posted Python answers in our sample data. Among these 16 381 users, 5419 unique users post accepted answers and 10 962 users post non-accepted answers. Table 6 shows a detailed distribution of the type of posts made by the unique users. Also, among all the answers posted, most of the users posted non-accepted answers.

Table 6: Distribution of Python question and answers among developers

Category	Count	Percentage
Users who post only questions	13 377	87.08%
Users who post questions and accepted answers	967	6.29%
Users who post questions and non accepted answers	1169	7.61%
Total users posting questions	15 361	
Users who post only accepted answers	4452	82.15%
Users who post accepted answer and questions	967	17.84%
Users who post accepted and non accepted answers	1484	27.38%
Total users posting accepted answers	5419	
Users who post only non accepted answers	8461	77.18%
Users who post non accepted answers and questions	1169	10.66%
Users who post non accepted and accepted answers	1484	13.53%
Total users posting non accepted answers	10 962	

Summary for RQ 1: We notice an increasing trend in the number of Python questions posted each year (except in the year 2021). We also notice a decreasing trend in the percentage of questions getting accepted answers. This finding shows a need for support for Python developers as the number of Python questions is increasing over the years. We notice a high satisfaction among the Python developers as only 0.37% of questions are unsuccessful and a large percentage of questions receive an accepted answer or first answer within the first hour after asking the question.

3.2. RQ 2: How did the Python tags evolve over the years?

Motivation: Tags are an essential part of Stack Overflow questions since they help organise them into distinct categories. Using tags to classify questions will also help to bring the attention of experts to each question[28]. Tags can also be used to filter the questions that we are interested in. In this RQ we aim to learn more about the various tags linked to Python questions. Analyzing the tags linked with questions will assist in determining the number of questions associated with each tag as well as the areas where Python developers experience the most difficulties.

Approach: This RQ was carried out in two stages. First, we extracted all of the tags associated with all of the retrieved Python questions and estimated the number of questions posted against each of them. This data was grouped yearly to determine how many questions were posted each year that use these tags. The top ten tags in this data were used to create a line chart to determine the trend of these tags over time.

Second, we used all tags identified as part of Section 2.2 and group them into different categories. Thirty questions were manually verified on the website by the first author for each of these tags in order to classify them into relevant groups. This was conducted to identify the complex and challenging areas of Python by determining the areas where developers are posting more questions. Table 7 provides a full list of the groups and the

[28]https://stackoverflow.help/en/articles/5611195-overview-of-tags

Table 7: Grouping detail of identified Python tags to extract questions

Group	Tags	% of Total Python Questions
Web development	django, django-models, cherrypy, django-queryset, flask, django-views, urllib2, django-orm, django-admin, gunicorn, django-shell, django-templates, flake8, django-media, django-authentication, django-signals, django-2.0, django--urls, django-manage.py, manage.py, django-migrations, django-rest-framework, django-custom-manager, django--aggregation, django-staticfiles, django-class-based-views, django-modeltranslation, django-q, django-1.10	17.8%
Scientific computing	pandas, pandas-loc, pandas-groupby, numpy-ndarray, numpy, python-imaging-library, beautifulsoup, scipy, pytorch, numpy--einsum, pyspark, scikit-learn, scrapy, pandas-explode, nltk, google-api-python-client, pyarrow, numpy-slicing, theano, fillna, h5py, pytables, tensorflow, keras, opencv, mplcursors	16%
General programming	python-module, python-class, python-datamodel, python--packaging, python-internals, python-c-api, python-typing, cpython, python-datetime, ironpython, timedelta, python--nonlocal, fnmatch, imaplib, itertools, pprint, pyperclip, smtplib, python-control, nonetype, python-dataclasses, pytz, difflib, python-attrs, python-closures, mod-python, python-click, jython, jpype, pypy, f-string, timeit	19.6%
OS, multithreading and multiprocessing	python-multithreading, python-os, python-multiprocessing, pyyaml, python-asyncio, popen, configparser, pathlib, django--celery, python-rq, argparse	6.75%
Installation, deployments and IDE	pip, python-import, pyenv, python-venv, pypi, setup.py, pyc, ipython, pythonpath, pywin32, python-idle, distutils, py2exe, python-poetry, pyinstaller, pydev, pycharm, jupyter--notebook, anaconda, jupyter, conda	12.9%
Testing and documentation	zen-of-python, pytest, python-unittest, pylint, pyflakes, pychecker, mypy, python-mock, python-logging, django--testing, django-validation, python-sphinx, imghdr, magicmock, python-docx, python-nose, pep8	10.4%
File handling and formats	python-unicode, python-wheel, python-zipfile, shutil, openpyxl, pypdf2, pypdf, xlrd	4.91%
Database and interaction	mysql-python, psycopg2, flask-sqlalchemy, pymongo	2.45%
Network and serialisationing	python-requests, pickle, pycurl, eventlet, paramiko, gevent	3.68%
Graphs and plotting	matplotlib, seaborn	1.23%
GUI support	wxpython, pygtk, tkinter, pyqt, pyqt4	3.07%
Others	pycrypto, pygame	1.23%

associated tags for each group. This data is used to plot a pie chart (i.e., Figure 7) to better comprehend the challenging areas for Python developers.

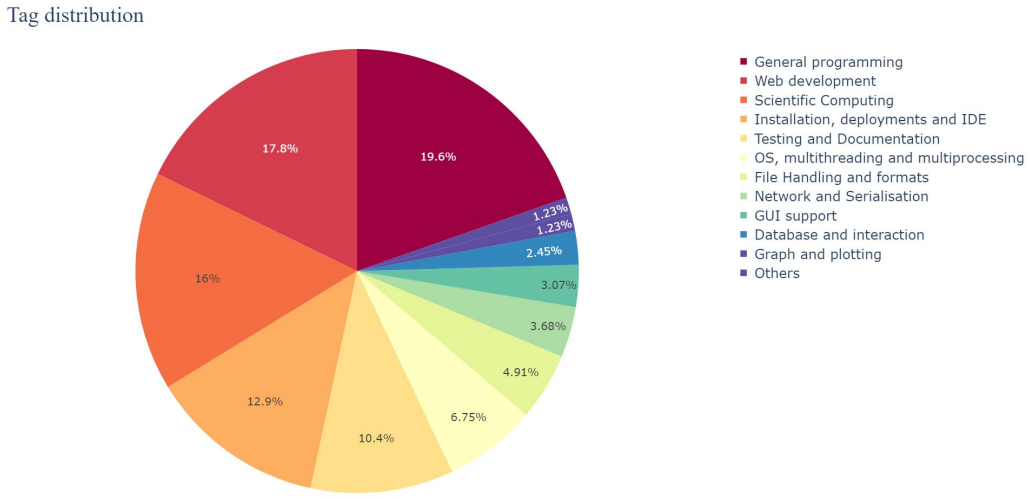

Figure 7: Tag distribution

Results: Our "Python-related" tag lists consist of 174 unique tags. We computed the total number of questions posted for each of these tags. The tag "Python" has the most questions submitted against it, with a value of 1 960 201 (80.02 per cent of all questions). Python-3.x, with a value of 314 799 (12.85%) appears the second most frequently in the dataset. This research focuses on Python and hence it is normal to have a high-frequency question consisting of Python tags. Here, we are more interested in identifying what other tags occur frequently in Python questions. Hence, we removed the tags like Python, Python-3.x, Python-2.x, Python-2.5, Python-2.7, Python-3.6, Python-2.6, Python-3.3, Python-3.5, Python-3.4, and Python-2to3 from our analysis (for this RQ), so that we can focus our analysis on other tags.

We computed the total number of questions posted for the rest of the tags and extracted the top 10 tags based on the count of the number of questions posted for each tag. We found that Django (291 137 posts or 11.88%), pandas (246 501 posts or 10.06%), NumPy (102 519 posts or 4.18%), dataframe (85 074 posts or 3.47%), TensorFlow (76 974 posts or 3.14%), list (68 603 posts or 2.8%), OpenCV (67 350 posts or 2.74%), matplotlib (65 393 posts or 2.66%), flask (50 403 posts or 2.05%), and dictionary (46 061 posts or 1.88%) are the top ten tags based on the number of questions. When examining the yearly top ten tags separately, it was discovered that these tags consistently appeared in the top ten spots for most of the years. The line chart in Figure 8 shows the yearly distribution of the number of questions posted against the top ten tags.

The graph shows a rapid rise in Django questions between 2008 and 2020, followed by a slight decrease from 2020 to 2021, which may be attributed to a drop in questions during that time as seen in the RQ 1. Tags such as list, OpenCV, dictionary, NumPy, matplotlib, and dataframe were steady from 2008 to 2010 and then showed an increase till 2021. However, from 2008 to 2010, tags like TensorFlow, Flask, and Pandas did not appear in the top 10. However, since 2010, there have been many more questions in pandas, and by 2020, they had surpassed the number of questions in Django. The study by McKinney [14] discusses in detail how "pandas" is one of the greatest tools for quickly handling and manipulating

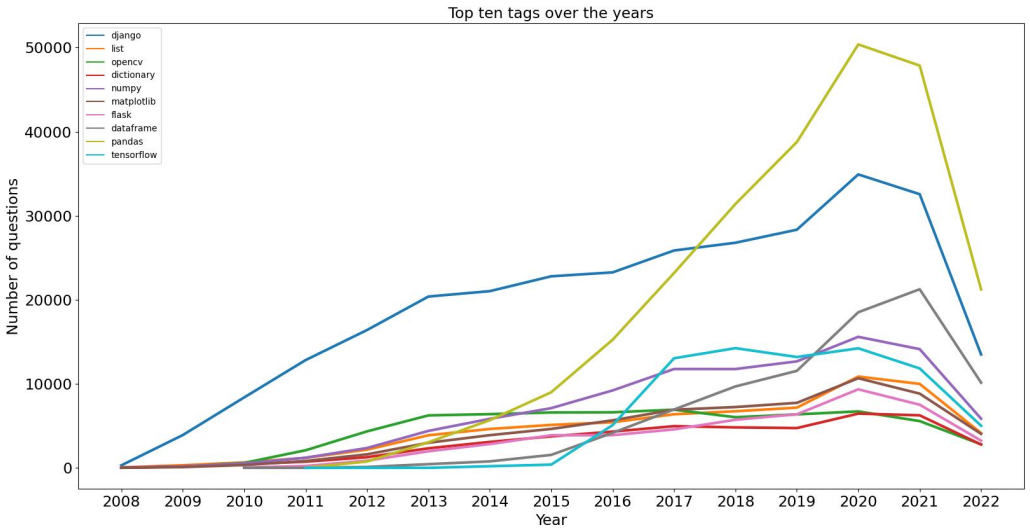

Figure 8: Top ten Python tags over the years

data, making it an essential tool for data analytics. The rise in popularity of data analytics and artificial intelligence since 2010 may be responsible for this spike in questions[29].

Questions on TensorFlow were consistent until 2011 when they suddenly increased in 2015, overtaking OpenCV, NumPy, and matplotlib to become the tag with the third-most questions posted in 2017. This may be the effect of the library being open-sourced by Google in 2015[30]. Additionally, the library is ranked first in the list of the most popular deep learning libraries in 2017 created by data incubators using data from GitHub, Stack Overflow, and Google searches[31]. Beginning in 2010, the top ten list also included the tag "flask", which continued to see a surge in question volume until 2020. The number of questions linked with the tag "dataframe" increased from 2017 to become the tag with the third most questions posted in 2021.

To ascertain which categories the tags belong to, we categorised them as shown in Table7. The majority of tags were linked to general programming (32 of 174 tags or 19.6%), and this included tags relating to data types (e.g., nonetype), classes (e.g., Python-dataclasses), time calculation and iteration tools (e.g., timedelta, itertools), and many more. Web development is the second category with the greatest amount of tags (29 of 174 tags, or 17.8%), with most of the tags being associated with the well-known Python web frameworks Django and Flask.

Scientific Computing is the third category (26 of 174 tags, or 16%), and it includes tags for computer vision and natural language processing (OpenCV, PyTorch, nltk), machine learning and artificial intelligence (scikit-learn), data manipulation and mathematical operations (pandas, NumPy). The next category, Installation, deployments, and IDE (21 of 174 tags or 12.9%), comprised tags for installation (pip, pyinstaller, etc.) and integrated development environments (IDEs), such as pycharm, jupyter-notebook, etc. The category Testing and Documentation (17 of 174 or 10.4%) featured tags for testing (such as pytest

[29]https://en.wikibooks.org/wiki/Data_Science:_An_Introduction/A_History_of_Data_Science#Chapter_Summary

[30]https://hub.packtpub.com/tensorflow-always-tops-machine-learning-artificial-intelligence-tool-surveys/

[31]https://www.thedataincubator.com/blog/2017/10/12/ranking-popular-deep-learning-libraries-for-data-science/

and Python-Unittest), error checking (such as Pyflakes and PyChecker), and documentation (such as Python-DocX and Zen-of-Python).

OS, multithreading, and multiprocessing tags (11 of 174 tags or 6.75%) include Python-asyncio, Python-os, etc., while File Handling and formats tags (8 of 174 tags or 4.91%) include pypdf, openpyxl, and others. Eventlet, Paramiko and other tags are categorized in the Network and Serialization category (6 of 174 tags, or 3.68%), whereas tkinter, pyqt, and other tags are found in the GUI Support category (5 of 174 tags, or 3.07%), and mysql-python, pymongo, and other tags are found in the Database and Interaction category (4 of 174 tags, or 2.45%). The least amount of tags (2 of 174 tags, or 1.23%) was found in the categories Graph and plotting with tags matplotlib and seaborn, and Others with the tags pycrypto and pygame.

Summary for RQ 2: The findings showed that the popularity of data analytics is rising indicated by the increase in the number of questions in the "pandas" tag. It was also observed that Python developers still struggle with general programming questions as most of the question tags belonged to this category.

3.3. RQ 3: What are some of the main topics of Python discussed by Python developers?

Motivation: Stack Overflow allows developers to use natural language to describe their challenges or propose a solution to problems. Every language has features and capabilities that are unique to it. For instance, Python developers must be proficient in a wide range of Python libraries, including Django, pandas, timedelta, and many others. As a result, depending on the features of each language and its applications, the issues that developers discuss may differ substantially. For this reason, it's critical to comprehend the main Python discussion points and the challenging areas that developers face daily.

In this RQ, we use the Stack Overflow dataset to identify the important Python topics that are commonly discussed by Python developers. Identifying the primary topics of discussion in the Python community will aid in determining the most interesting areas for Python developers. This will contribute to the enhancement of Python's tools, design, and documentation. The information from this RQ will also be valuable to Python-related publications because it identifies areas in which the Python community is most interested and has many issues. Researchers can use this to find possible study areas in the future.
Approach: To answer this RQ, we use **topic modelling** to determine the major Python discussion topics. Using the word distributions in the document collection, topic modelling is an unsupervised machine learning algorithm, that attempts to find any hidden patterns of word frequency. Based on the patterns, it provides a collection of topics that contain collections of words that co-occur in the document set [15].

We use the following methods to identify the main topics discussed by the Python community. **First**, we extract the entire Python posts from Stack Overflow and use several pre-processing techniques to extract relevant keywords. **Second**, topic modelling approaches are used to extract the topics from the pre-processed data. **Third**, to learn more about the context around the discovered topics, the extracted topics and a sample of questions are analysed, and a name is given to it based on the context. Figure 9 shows the approach followed for topic modelling. Following is the detailed description of each step.

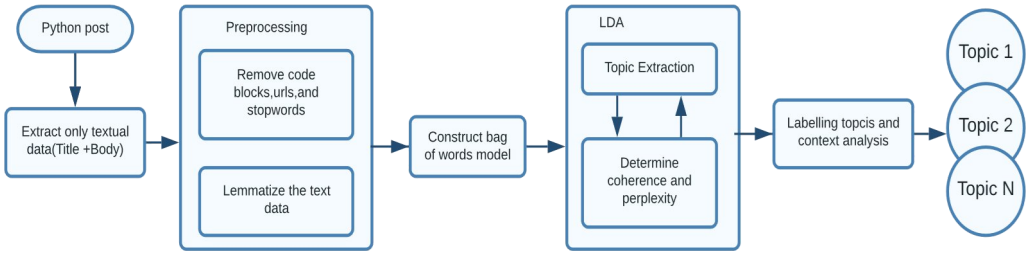

Figure 9: Topic modelling approach

1. Data Extraction and Pre-processing: The Stack Overflow dataset has a large number of questions and hence, we created a random sample of question posts with a 99% confidence level and a 1% confidence interval. We extracted 16 529 question posts. We extracted both title and body of these questions for analysis. In the initial round of pre-processing, we eliminate all code snippets from the text. As the code blocks do not add meaningful words for topic modelling, this step is critical to remove extraneous noise from the data. We exclude all URLs and HTML tags from the textual data because these are not required to generate keywords.

The next step is to eliminate any punctuation and expand any word contractions (for example, a change I'm to I am). Multiple white spaces, numerals, and words that are less than three characters in length are then eliminated from the text. Then we exclude the English stop words offered by NLTK as well as a few custom stop words like "thanks", "question", "answer", etc. Then, to assist with topic modelling, the words are lemmatized to their root form using the NLP Python module spaCy [16] Then, a bag of word models is constructed for Topic modelling to efficiently determine the word occurrences in all of the questions [17].

2. Topic Modelling: The topic model is a statistical model that is commonly used in natural language processing to determine the primary topics in a document collection[32]. A popular algorithm for topic modelling Latent Dirichlet Allocation (LDA) [18] is used to answer this research question. LDA is a probabilistic model created from a set of documents, where each document is represented as a mix of topics, and each topic is characterized by the word distribution. Using LDA for topic modelling can be seen in multiple prior research which use Stack Overflow dataset in a similar context and has proved to produce effective results [3, 8, 13, 19, 20].

The LDA algorithm receives the corpus prepared using the bag-of-words technique. The number of topics, or "k", is another key parameter needed by the LDA algorithm. Determining an ideal k value is critical in LDA since a low number provides a broad selection of topics while a high value provides more detailed topics that are essentially noise. To determine the ideal k value, we considered a parameter called topic coherence. The semantic relationship between the terms in a topic is measured by the topic's coherence, and a greater coherence value produces better results [21]. So, we ran the LDA iteratively with the value of k ranging from 2 to 100 and with a step size of two. The coherence values for each cycle were examined, to find the ideal k value. We identified k value 20 as giving the best results.

3. Determining the topics: After the model with the optimal coherence score has been determined, research is conducted to ascertain each LDA topic's nature and meaning. As each word in the corpus is given a probability of belonging to a certain topic by the

[32]https://en.wikipedia.org/wiki/Topic_model

LDA method, the distribution of topics across words is employed for topic labelling. To comprehend the topic, the top word with the highest probability is utilized. Each document is also assigned a probability that indicates how closely it relates to the topic. A sample of these documents was manually identified to better support the topic determination. These techniques helped in assigning an appropriate label to each topic.

Results: The LDA algorithm does not provide meaningful topic names for the results that it generates. Hence, two of the authors perform a manual examination of these 20 topics and associated keywords to find meaningful topic names. For each topic, we looked at the terms that are present in the topic and performed a discussion to assign a meaningful name. During this analysis, we notice that several topics contained similar words with a high probability, thus they were merged to better represent the topic. Using this process, we finally left with nine topics: *general programming, scientific computing, web development, file handling and formats, data structures and formats, graph and plotting, operating systems, multi-threading and multiprocessing, installation, IDE and deployments, and GUI support.* The nine topics, the number of documents that correspond to each topic, and the top keywords linked with each topic are all displayed in Table 8.

Examining the top words in each topic was beneficial in identifying the topic name. However, more in-depth analysis is needed to determine the rationale or the problems

Table 8: Topics generated by LDA with the frequency of occurrence and relevant keywords

Topic	Question count	Percentage %	Top keywords
General programming	3826	23.14%	input, output, time, differ, result, length, function, binari, random, variabl, random, class, local, default, method, argument, multipl, actual, refer, check
Scientific computing	2602	15.74%	panda, datafram, model, test, train, dataset, tensorflow, keras, predict, fit, featur, pyspark, spark, accuraci, scrapi
Web development	2478	14.99%	django, page, form, view, html, server, web, flask, post, applic, load, http, site, app, respon
Data Structures and Formats	2421	14.64%	arrai, list, vector, object, float, integ, data, iter, json, text, dictionari, tupl, dict, type
Installation, deployments, and IDE	1396	8.44%	instal, import, modul, version, packag, pip, path, environ, librari pycharm, configur
OS, multithreading, and multiprocessing	1172	7.09%	process, start, task, script, command, job, thread, execut, port, worker, client, socket
Graphs and plotting	1099	6.64%	imag, plot, matplotlib, background, color, label, figur, displai
File Handling and formats	922	5.57%	file, line, open, csv, format, directori, zip
GUI Support	613	3.7%	item, button, click, window, tkinter, press, option, page, menu

encountered by the developers. Hence, two of the authors perform a manual analysis of a statistically significant set of questions. We use a confidence level of 95% and an interval of 5%. Using these parameters, we created a sample size of 376 posts. Two of the authors annotated the dataset with their understanding of the rationale behind each topic. Once, completed both of the authors exchange the annotated dataset with each other. During the review, the conflict was resolved using a discussion between the annotator and the reviewer. There were 17 posts for which no agreement was received. Hence, an agreement percentage of 95.2% is achieved.

3.3.1. General programming

This topic covers queries about general programming, for example, Object-Oriented (OO) principles, using Python with other programming languages, a general queries about implementing logic. Table 9 shows some example questions relating to this topic. One of the major concerns that we noted in this topic was implementing OO principles in Python. For example, question 1 (QId: 65 676 114) in Table 9 shows a query where the developer is asking a question about "importing class from different scripts'. Importing classes or global variables from different files requires an understanding of the OO concepts, an incorrect initialization will lead to an error. The second main query in this was about using Python with other programming languages, for example, javascript, ruby, C++, SQL, etc. question 2 (QId: 26 071 943) in Table 9 shows a query where the developer is asking a question about how to use Ruby and Python together. This reveals that even though many individuals are familiar with Python programming concepts, creating practical applications that integrate with libraries of other languages is still challenging. Developers use Python in various applications and hence several times they need to use Python with other programming languages. Providing appropriate support to the developers when they are using Python with other programming languages can be useful. We also observe several queries related to the logic of code with string, lists dictionaries, time-date operation, etc. Question 3 (QId: 21 162 471) in Table 9 shows an example of a query where a developer seeking help with reversing a string.

Table 9: Example of question belonging to the topic "General Programming"

S. No.	QId	Title
1	65 676 114	How to **import class** from different script, where the class uses a global variable at initialization?
2	26 071 943	control stdin and stdout of a **ruby** program in python
3	21 162 471	**Search and replace string** with reverse

In summary, from analysis of this topic, we find that developers face difficulty in applying and using OO principles, using Python with other programming languages, and implementing logic with data structures like strings, lists, dictionaries, etc. To help developers with these more tutorials can be written to use the OO principle in Python. Developers need support in terms of API when they are integrating Python with other programming languages. Hence, the current API can be improved so that it becomes easier for developers to use Python with other programming languages. This is also the **biggest topic** if we see the size of the topic, i.e., the number of questions in each topic, and hence by addressing queries related to this topic needs of a large number of developers can be addressed.

3.3.2. Scientific computing

This topic covers queries about advanced computing concepts that solve complex challenges, such as artificial intelligence, machine learning, neural networks, computer vision, etc. Most of the conversations in this area concerned handling data in the proper format for implementing various models, big data libraries, and general data science concepts. Some examples of questions belonging to this category are shown in Table 10. The majority of the inquiries were on how to efficiently handle and operate data using Panda's library. For example, in example 1 (QId: 41 814 828) in Table 10, the developer is seeking help with reading nested JSON in Pandas data frames. In addition to the pandas' package, there are numerous more queries about data preparation before model implementation. For example, refer to example 3 (QId: 65 255 029) in Table 10 where the developer is asking a query about converting data to glove or word2vec format. A lot of developers asked for help utilizing PySpark to operate on data, such as in example 4 (QId:70 520 386) in Table 10, the developer is asking a question about extracting a specific column in Pyspark. Another prominent topic of discussion in this field is the theories of many machine learning models, algorithms, performance measurements, and scientific computing concepts. For example, consider example 5 (QId: 59 793 818) in Table 10, where the developer is seeking help in understanding the negative cross-validation score function. We also notice some questions related to data crawling (QId: 34 542 381).

Table 10: Example of question belonging to the topic "Scientific Computing"

S. No.	QId	Title
1	41 814 828	Trouble reading nested JSON in **pandas** dateframe
2	49 517 830	Incorrect regex identification using **pandas**
3	65 255 029	How can I convert a **dataset** to glove or word2vec format?
4	70 520 386	Extract specified string from a column in **PySpark dataframe**
5	59 793 818	What does negative **cross_val_score()** mean?

In summary, from this topic, we observe that developers face issues in data extraction and data manipulation when using libraries such as Pandas and pyspark. The research community should provide tools to the developer community so that they can efficiently and easily process data. Joy et al. [22] performed a detailed study about questions related to Pandas on Stack Overflow and identified five major categories in which developers face difficulty. Among them, the data frame-based issues are at the top position which matches the finding of our paper. They also need help in understanding the output of machine learning libraries. Hence, these libraries can be improved to help developers in interpreting the results. For example, if a model is outputting negative cross-validation then the developer can be informed that there is some error in the code and the developer needs to correct it.

3.3.3. Web development

This topic primarily discusses various aspects of web frameworks and the challenges that developers encounter in web development. The two primary Python web frameworks, Django and Flask, were the major topics of discussion. Table 11 contains a few examples of questions in this category. The majority of Django discussions focused on user profiles,

forms, and models, among other topics. For instance, example 1 (Question Id: 12 526 065) shows a new Django developer who needs assistance in setting up a user profile for a web application. While seeking help on Stack Overflow, the developer also criticizes the lack of documentation for this concept.

Table 11: Example of question belonging to the topic "Web Development"

S. No.	QId	Title
1	12 526 065	**Django:** create new user and profile
2	52 264 764	how to access field using **django** signals
2	53 243 629	How to retrieve logged user id in **flask**

Additionally, a developer in example 2 is having trouble using Django signals to access fields. The developer claims that the documentation was insufficient to fix the problem. Also, a lot of developers had queries about the Flask framework. For example, a developer is requesting information about a flask library to retrieve the user id after user login in Example 3 (53 243 629) of Table 11.

In conclusion, we can see from this topic that developers struggle to understand ideas from existing documentation, which needs to be improved. An alternative approach might involve adding more courses that focus on building web applications from start using Django and Flask by providing in-depth examples that can assist developers.

3.3.4. Data structures and formats

This topic covers queries about operations on data structures like lists, dictionaries, arrays, strings, vectors, etc. In this topic, we noted three main developer concerns, 1) efficient use of these data structures, 2) conversion from one data structure type to another type, and 3) using nested data structures. Table 12 shows some examples of questions that belong to this topic. Question 1 (QId: 5 324 217), in Table 12 shows an example where the developer is asking about the processing of nested lists. Question 3 in this table, shows an example of queries where developers are seeking advice about turning the list into a dictionary. In large programs, it often happens that developers need to convert from one data structure to another hence simple functions that can help developers in making such conversions can be useful. In question 4 (QId: 25 982 638), the developer is asking a question about finding, a fast way to convert a list to an array.

Table 12: Example of question belonging to the topic "Data Structure and Formats"

S. No.	QId	Title
1	53 242 179	Comparing multiple **lists inside of lists**
2	39 663 866	Manipulating **dictionaries within lists**
3	58 340 665	**Turning list into dictionary**
3	25 982 638	H5py: **fast way** to save list of list of numpy.ndarray?

In summary, from analysis of this topic, we observe that the developer faces difficulty using nested data structures and converting from one data structure to another. The research community and the Python developer community need to design some methods or APIs that make such kinds of operations easier. In addition, we observe that the developers need help

in optimizing their code. Code optimization is an important problem and it is noticed that developer face difficulty in it [8]. Permua et al. [8], also reported that developers face difficulty in code optimization. They reported issues related to code optimization and refactoring such as simplifying switch-case statements or loops, duplicate code removal, loops, compacting logic, etc. Our finding provides another dimension to these results by suggesting that developers seek help in code optimization when converting from one data structure to another.

3.3.5. Installation, deployments, and IDE

This topic covers issues related to installation, import errors, deployments, and problems with Integrated Development Environments (IDEs). We notice three main concerns in this topic, 1) issues with the Virtual environment, 2) resource access or import error, and, 3) issues with pip/conda install. Table 13 shows a few examples of questions related to this topic. The first major issue that we identified was related to Python *virtual environment*. A virtual environment is created on top of the "base" Python environment. It is used to isolate from the package environment of the base environment. We noticed two major concerns of the developers first, issues with package installation in the virtual environment and the wrong listing of packages in the virtual environment. Question 1 (QId: 41 801 382) shows a query where the developer is asking questions about installing packages in a specific virtual environment. The second major concern was resource access or import error, i.e., the developer was facing issues related to access to a certain resource especially when they integrated Python with other libraries/technologies such as docker, flask, AWS, google app engine, cython, WIN API, psycopg2, etc. Question 2 (QId: 45 266 200), shows an example question where the developer is asking for help to access a resource in Python when it is integrated with AWS. The third main concern we notice was related to the installation of packages using pip or conda. We notice that developers face lots of difficulty in installing packages because of different versions of Python and pip or different operating systems (e.g., Linux, windows, etc.). Table 13 shows an example (question 3, QId: 30 993 086) where the developer is getting an error about the installation of pip.

Table 13: Example of question belonging to the topic "Installation, deployments, and IDE"

S. No.	QId	Title
1.	41 801 382	How to install packages into **specific virtualenv** created by conda
2.	45 266 200	Python in AWS Lambda: "module "requests" has no attribute "get"
3.	30 993 086	pip3: command not found but python3-pip is already installed

In summary, from the analysis of this topic, we notice issues related to the installation of packages and libraries. The research community and the Python development community need to provide some easy methods so that developers can easily install packages without the need to worry about the exact version. Because from the analysis, it is observed that the Python packages are not very stable, and using a different version can give errors. Hence, there is a need to make Python more robust by making it backwards-compatible.

3.3.6. OS, multi-threading, and multiprocessing

This topic covers discussion about OS process execution, subprocess management, job scheduling, client-server issues, database/SQL issues, and multitasking/multithreading.

We observed that the major concerns in this topic are 1) Operation with the "celery" framework, 2) Execution of threads, 3) Running multiple processes/subprocesses, 4) Database/SQL-related issues, and 5) Job scheduling. Table 14 shows some example questions in this topic category. The first major area of discussion in this topic category was on celery framework. Celery is an open-source asynchronous task queue or job queue which is based on distributed message passing[33]. An example is Question Id: 61 221 609 in Table 14 which discusses techniques for running many processes simultaneously. Similarly, Question Id: 53 044 978 is also on the Celery framework which has no answers. These questions do not have a satisfactory response, which could be an indication of a lack of Celery experts in the Stack Overflow platform.

Table 14: Example of question belonging to the topic "OS, multi-threading, and multiprocessing"

S. No.	QId	Title
1	61 221 609	How to get multiple **celery** task results at the same time?
2	53 044 978	How to use a **Celery** worker
3	32 565 819	Python socket programming with **threads**
4	10 604 958	Run and get output from a **background process**
5	13 060 427	sorting and selecting **data**
6	41 604 289	urllib2 <urlopen error [Errno 61] **Connection refused**>

Another major area of discussion within this topic is the execution of processes. For example, Question Id: 10 604 958 discusses strategies for getting output from background processes, and the developer also edited the post later to indicate that the responses offered were insufficient to resolve the issue. We also noticed many questions related to thread execution, an example is Question Id: 32 565 819 about socket programming with Python threads. This question also did not receive an accepted answer. Additionally, we observed that there are numerous questions regarding the scheduling of programs and execution of jobs at various times. We notice client server-related issues in this topic where the developer asks for issues related to making connections between client servers or processing/managing data between these two. Question Id: 41 604 289 shows, an example where the developer is getting a connection refused error when trying to connect to the server. We also notice several SQL/database-related issues, for example, Question Id: 13 060 427, shows a query from the developer where she is asking how to use Python and SQL together.

In summary, from analysis of this topic, we notice issues related to distributed systems, threads, processes/subprocesses, Database/SQL-related, and Job scheduling. We also notice a lack of satisfactory response in this topic area which indicates the need for more community experts to answer OS-related queries. The research community needs to work and provide an appropriate tool to the developers that make it easier for the Python developer such advanced operations.

3.3.7. Graphs and plotting

This section includes discussions about plotting and adding colours to graphs and it covers 6.64% of the total number of documents in the analysis. The major areas of discussion we observed in this category were on libraries like "matplotlib" and "plotly". Additionally,

[33]https://docs.celeryq.dev/en/stable/getting-started/introduction.html

there were a few questions about the "seaborn" library. Table 15 gives an overview of a few questions in this topic category.

Queries on the matplotlib library were most commonly seen within this topic. One example is Question Id: 22 923 402 in Table 15 where the developer is struggling with enlarging the axis size. Enlarging the size of the axis, labels and different colour schemes is crucial in graph plotting. Another example is Question Id: 64 568 227 in Table 15 where a developer struggles with adding grids to the background of a plot. This question did not receive an accepted answer.

Another major area of discussion in this category was the plotly library. For instance, consider the question with Id:64 949 228 (Table: 15), where a developer seeks help to draw a horizontal line in the graph in a different coordinate position. The developer also gives a sample output obtained for the same problem with the matplotlib library. This could indicate that many developers struggle with some basic operations in plotly, and improving the documentation with more examples could help to address this issue. Another example is Question Id: 70 889 046 (Table 15) where the developer asks about colouring Sunburst rings based on different conditions. The question did not receive any answers suggesting a lack of experts in the Stack Overflow platform.

Table 15: Example of question belonging to the topic "Graphs and plotting"

S. No.	QId	Title
1	22 923 402	**matplotlib:** enlarge axis-scale label
2	64 568 227	How to add a grid graph as a **background** of one graph plot?
3	64 949 228	Plot horizontal lines in **plotly**
4	70 889 046	How do I colour Sunburst rings in **Plotly** based on different conditions

In summary, from our analysis of this topic, we found that developers face difficulties in some of the basic concepts of axis size, labels and colour schemes in both plotly and matplotlib libraries. Also, we noticed a lack of experts in the Stack Overflow platform to provide answers regarding the plotly library. Improving the documentation on these libraries with more examples will make it easier for developers to understand the fundamentals of plotting. Also, we observed a lack of adequate community experts to answer questions related to the plotly library.

3.3.8. File handling and formats

This topic addresses queries pertaining to various file operations and the handling of multiple files. We noticed that some of the common areas of discussion in this category are managing multiple files simultaneously, errors in file handling, file operations like read, write, compress, etc., and altering the content of files. Some examples of questions belonging to this category are given in Table 16.

The majority of the questions were about operating with different file formats. Consider the example Question Id: 45 710 477 in the Table 16 where a developer is trying to use values from one Python file to another. There are many similar questions on Stack Overflow where developers struggle with handling two or more files. Another major area of discussion is regarding errors while handling files. One such example is 2 (Question Id: 62 133 256) in Table 16. Here, the developer gets an Out of Memory error while operating on a large file and seeks help on Stack Overflow. But the question did not receive any response from

Table 16: Example of question belonging to the topic "File Handling and Formats"

S. No.	QId	Title
1	45 710 477	Importing variables from **another file** in Python
2	62 133 256	numpy loadtxt for **large text files**
3	19 277 816	python logging: how to get **compressed** .gz log file using TimedRotating-FileHandler
4	55 481 848	How to **modify the contents** of text file?

the platform. Often, the search for and taking of corrective measures to correct errors in software development is extremely time-consuming and has a significant effect on the cost. This is explained in detail in the study [23].

The other major area of discussion was on compressing files, as in Example 3 (Question Id: 19 277 816) in Table 16 where a developer wants to compress the output file to a particular format. The question did not receive any answers from the platform. Another area of discussion was altering the contents of the file. For instance, example 4 (Question Id: 55 481 848) in Table 16 is about replacing commas in a file where a file seek operation can resolve the issue.

In conclusion, from our analysis of this topic, we noticed issues related to multiple file management, errors in file handling, file operations like read, write, and compression, and modification of file content. We could also observe a lack of response from the Stack Overflow platform in many of these questions, which suggests that more experts are needed to answer queries. Also, errors in file operations need to be addressed to reduce the impact on software development costs and time. So, the developer community can work on documenting some of the common errors and mitigation steps, especially in the case of large files.

3.3.9. GUI Support

This was the topic category with the fewest documents (i.e., 3.7%) and includes issues related to the graphical user interface (GUI). Most of the questions were on popular Python GUI libraries such as "pyqt", "tkinter", and "wxpython". Table 17 contains a few examples of questions belonging to this category. Considering the library "tkinter", an example question is 1 (Question Id: 54 122 578) in Table 17 where a developer struggles with labels in the window displayed. In example 2 (54 122 578) of Table 17 a developer keeps running into errors. The question did not receive any accepted answer. Another major discussion area was regarding the "pyqt" library. For instance, consider example 3 (Question Id: 22 299 612) from Table 17 where a developer encounters trouble displaying colour icons in the menu.

Table 17: Example of question belonging to the topic "GUI Support"

S. No.	QId	Title
1	54 122 578	**Tkinter** labels not showing in the pop-up window
2	54 122 578	**Tkinter** variable classes not defined
3	22 299 612	How to display colour icons in the menu, **PyQt**

We observed that many developers struggle with the basics of graphical user interface development. Although these questions received satisfactory responses from the platform,

measures can be taken by the educator community to create more courses in Python GUI libraries.

Summary for RQ 3: Our analysis of the Python posts reveals that there are nine major topics: *General programming, Web development, Scientific Computing, Data Structures and Formats, Installation deployments IDE, OS multi-threading and multiprocessing, Graphs and plotting, File Handling and Formats*, and *GUI Support*. Each of the topics has some main issues, however, we notice some major concerns like integrating Python with tools and languages, problems associated with understanding and changing between different data structures, issues related to Django and Flask, issues in installing packages using Pip/conda, issues related to the subprocess, multi-threading, etc.

3.4. RQ 4: What are the most popular and challenging topics discussed by developers in Stack Overflow?

Motivation: In this RQ we aim to identify the type of Python questions on Stack Overflow that developers find interesting and challenging. This RQ is based on the results from RQ 4, and the analysis is carried out in a more fine-grained manner to comprehend the Python topics that are thought to be the most well-liked and difficult by the Python community. This will be useful to obtain more detailed and focused insights on the patterns of Python questions, problems, and demands on supporting activities. Additionally, identifying the popular and difficult topics might enable organizations like Kaggle and Stack Overflow to assign and hire additional experts in those areas.

Approach: As in earlier research, we utilise three complementary popularity measurements to determine a topic's popularity [8, 13, 24–27]. First, the average number of registered and unregistered users who have viewed the post. More views suggest that more developers have encountered similar difficulties, hence the topic is more popular. Overall, this metric measures the Python community's interest based on how frequently the post is viewed.

Second, the average number of posts identified as favourites by the community. The favourite count measure illustrates how useful the question was. Additionally, the higher favourite count suggests that more individuals have encountered the same problems and found the solution useful. Overall, this metric assesses the popularity of questions depending on how useful they are.

Third is the average score for each of the posts. On Stack Overflow, users can give a question an upvote if they think it is useful. The score is then calculated using these upvotes. A higher score indicates that more developers find the subject intriguing and have faced similar problems. Therefore, we utilise this as a measure of each question's perceived community value.

As in previous research, two metrics are employed to determine the difficulty of each topic [8, 28]. First, the number or percentage of posts without an accepted answer is evaluated. Only the person who asked the question has the capability to mark an answer as an accepted answer when they are satisfied with the solution. Therefore, a question with an accepted answer is an indication of user satisfaction. Also, the fact that so many questions have been made without any accepted answers shows how difficult the question in the topic is to answer [13, 29]. Second, we look at the question in a topic for which there are no answers. Sometimes, even when they obtain a suitable response, users who ask

questions neglect to designate the response as accepted. Therefore, considering the queries for which there is no answer, it may be inferred that neither experts nor developers were able to provide an answer, making the question difficult [8, 28].

Results: The results are subdivided into two sections topic popularity and topic difficulty.

Topic Popularity: Table 18 shows that the most popular topics are Data structures and Formats and Installation, Deployments, and IDEs, while File Handling and Formats are the least popular. Even though there are fewer questions related to Installation, deployments, and IDE than General programming, the average view count for the former is larger than the latter. This implies that many developers are constantly seeking and viewing the installation and IDE-related questions. The most popular topic is Data structures and Formats, according to the other two metrics average score and average favourite count. This demonstrates that the solution to issues based on data structures was more helpful and satisfactory to the developers.

Table 18: Popularity of Python topics

Topic	Count	Average view count	Average score	Average favorite count
General programming	3826	2082.792	2.048	0.605
Scientific computing	2602	1883.971	1.640	0.500
Web development	2478	2057.166	1.982	0.499
Data Structures and Formats	2421	6082.245	4.495	0.980
Installation, deployments, and IDE	1396	5707.434	4.590	1.113
OS, multithreading, and multiprocessing	1172	1964.17	1.732	0.479
Graphs and plotting	1099	2253.111	1.626	0.501
File Handling and formats	922	1785.682	1.085	0.284
GUI Support	613	1908.889	1.334	0.368

The question with the highest score and favourite count in the entire dataset is "What does the 'yield' keyword do?" with a score of 10 150 and 6431 users selecting it as their favourite post, evidences the lack of documentation. It is interesting to note that the findings of this study suggest that questions about Installation, deployments, IDE, and Data structures and formats provide the most challenges to Python programmers, as illustrated by the higher view and favourite counts of these topics. However, questions about Data Structures and Formats receive more satisfactory answers from the experts on the Stack Overflow platform. The rising popularity of these issues can be addressed by creating documentation on every functionality and library that are more comprehensive and accessible.

Topic Difficulty: Table 19 shows the difficulty level of various topics. We categorize the topics into four levels of difficulty using the metric percentage of questions without answers (PQ-WAnA) and percentage of questions without an accepted answer (PQ-WAcA). Following are the levels we created (level 1 is the most difficult and level 4 is the least difficult):

level 1 = [PQ-WAnA $\geq$ 0.229 and PQ-WAcA $\geq$ 0.621],
level 2 = [0.180 $\geq$ PQ-WAnA $\leq$ 0.228 and 0.529 $\geq$ PQ-WAcA $\leq$ 0.620],
level 3 = [0.162 $\geq$ PQ-WAnA $\leq$ 0.179 and 0.482 $\geq$ PQ-WAcA $\leq$ 0.528],
level 4 = [PQ-WAnA $\leq$ 0.162 and PQ-WAcA $\leq$ 0.481].

Table 19: Difficulty of Python topics

Level	Topic	Count	Without accepted answer %	Without answer %
3	General programming	3826	0.481%	0.162%
3	Scientific computing	2602	0.461%	0.1575%
3	Web development	2478	0.515%	0.186%
4	Data Structures and Formats	2421	0.422%	0.111%
1	Installation, deployments, and IDE	1396	0.621%	0.229%
2	OS, multithreading, and multiprocessing	1172	0.534%	0.189%
2	Graphs and plotting	1099	0.529%	0.224%
3	File Handling and formats	922	0.517%	0.162%
3	GUI Support	613	0.513%	0.197%

Table 19 reveals that the Installation, Deployment, and IDE issue has the most questions without any accepted answers or answers. We observed that questions related to Installation, IDE and deployment were also in the popular topic category. This implies that a large number of individuals are looking for these questions and not getting satisfying answers. The lack of answers may be caused by poorly constructed or ambiguous questions that lack specific details about the issue at hand. This, however, emphasizes the significance of the need for efficient resources and support on package installation and IDE-related issues in Python.

To avoid the difficulties faced by developers in the early stages of development, it is essential to address these issues. Another noteworthy finding is that the issue of data structures and formats is both popular and less complex. This topic has the lowest percentage of questions without accepted answers and answers. This can suggest that, even when developers encounter data structure and format-related concerns, the community can supply adequate and acceptable answers.

Correlation between topic popularity and difficulty: Similar to the previous studies [8], we performed a correlation analysis between the popularity and difficulty metrics. We compute the Pearson correlation between six pairs of metrics, 1) average view count and the percentage of questions without accepted answers, 2) average view count and the percentage of questions without answers, 3) average Score and the percentage of questions without accepted answers, 4) average score and the percentage of questions without answer, 5) average favorite count and the percentage of questions without accepted answers, and 6) average favorite count and the percentage of questions without answer. Figure 10, shows the scatter plots for all six pairs. This figure shows a weak correlation between popularity and the "percentage of the question without any answer" difficulty metrics. We do not observe much correlation between popularity metrics and the other difficulty metrics (i.e., percentage of questions without accepted answers). This figure shows two outlier topics, i.e., "installation deployment and IDE" topics and "Data structures and formats topics. The installation deployment and IDE topic has high popularity and difficulty whereas the "Data structures and formats" has high popularity but low difficulty, i.e., this topic is popular but receiving satisfactory answers from the community.

To obtain a deeper analysis of the correlation, we analyze Pearson and Spearman correlation coefficient between popularity and difficulty metrics. Table 20 shows the Pearson correlation value obtained for each of the pairs. We obtained a positive correlation between % of questions without accepted answers and all the three popularity metrics whereas

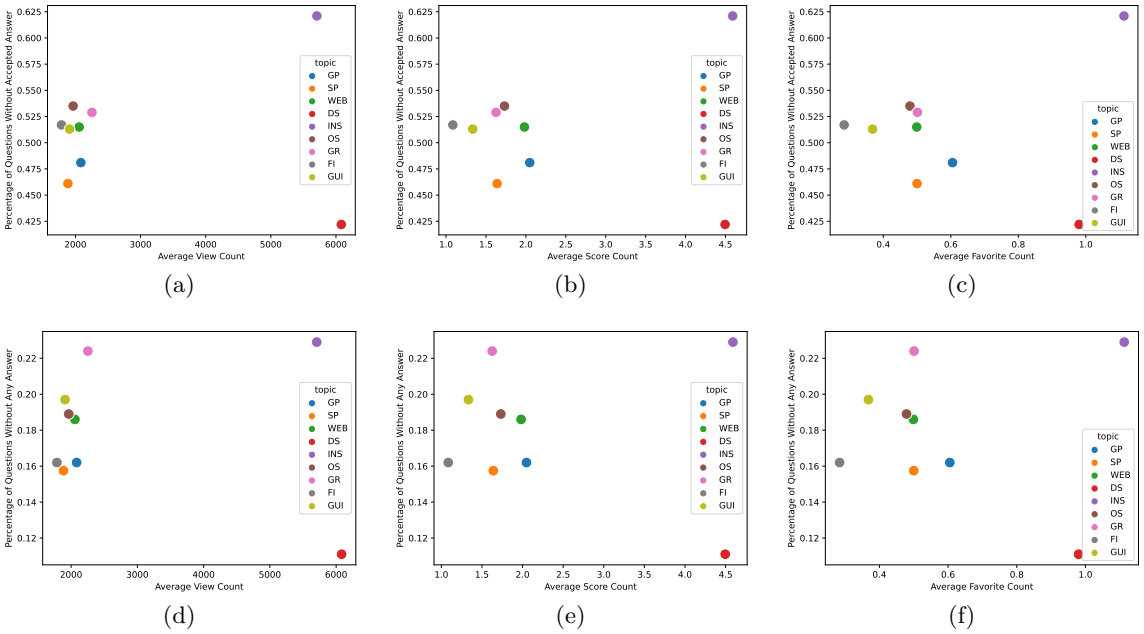

Figure 10: Scatter plots showing the correlation between popularity and difficulty matrix for all the nine topics: GP: General Programming, SP: Scientific Computing, WEB: Web Development, DS: Data Structure, INS: Installation Deployment and IDE, OS: OS, Multi-threading and multiprocessing, GR: Graphs and Plotting, FI: File handling and Formats, GUI: GUI support

a negative correlation between % questions without answers and all the three popularity metrics. However, none of this correlation is statistically significant as indicated by the p-values. Next, we perform Spearman correlation analysis. Table 21 shows that there is a low positive correlation between the average view count and percentage of questions without any answer using Spearman correlation. However, this is not statistically significant as indicated by the p-values. To further explore the reason *why some questions are not getting any answers*, we manually analyzed a sample of unanswered questions.

Table 20: Pearson Correlation between Popularity and Difficulty Metrics, Cor: correlation

	Average view count (cor/p-value)	Average score (cor/p-value)	Average favorite count (cor/p-value)
% w/o any answer	−0.164/0.671	−0.133/0.732	−0.04/0.91
% w/o accepted answer	0.071/0.855	0.104/0.788	0.164/0.67

Analysis of questions without any answer: In the next step, we analyzed questions without any answer for the topic "*installation, deployment, and IDE*". We extracted a statistically significant sample of 172 questions. This sample represents a confidence level of 95% and an interval of 5% from a total of 308 unanswered questions from this topic. Two of the authors manually analyzed these questions. From our analysis, we observed that there were very few questions that did not get any comments from the Stack Overflow community. Those consist of low-quality questions or questions which needed more details.

Table 21: Spearman Coefficient between Popularity and Difficulty Metrics, Cor: correlation

	Average siew sount (cor/p-value)	Average score (cor/p-value)	Average favorite count (cor/p-value)
% w/o any answer	0.175/0.651	−0.033/0.93	0.033/0.93
% w/o accepted answer	0.033/0.932	−0.033/0.93	−0.083/0.83

For most of the questions, the Stack Overflow community provided help in the form of comments where they discussed potential solutions, providing links to other related Stack Overflow questions, or other useful links (e.g., bug issues reports on GitHub, etc.). We notice in that there were many questions where the users were trying to integrate Python with other technologies or IDEs like C/C++, Cython, Pycharm, Spyder, AWS, and Docker, and hence were facing installation issues. We also notice several questions where users were facing issues in installing Python packages on the Linux environment.

Summary for RQ 4: Our analysis reveals that *Data structures and formats* and *Installation, Deployments, and IDEs* are among the most popular topics. *Installation, Deployment, and IDE* was found to be the most difficult topic. Our analysis of unanswered questions reveals that many unanswered questions are not difficult but they did have useful comments and some of them were not clear.

4. Discussions and takeaways

According to the findings of the study, Stack Overflow is an extremely popular forum for developers to ask questions and engage in discussions about the difficulties they face when programming in Python. In most circumstances, developers can ask for help in a variety of Python libraries and applications and receive a suitable response. To further our understanding of the difficulties faced by Python developers, this quantitative research is supplemented with a manual analysis of a statistically significant sample of posts. The results also assist educators in updating their course content to accommodate real-world situations and practical language applications. In this section, we provide details of findings that can help the community.

4.1. Research community takeaways

The findings of this study demonstrate the difficulties that Python developers face in real-world projects. These findings highlight the gap between the academic elements of Python and its practical real-world application. This provides an opportunity for the research community to better investigate and contribute to this field.

Major topics: According to the findings of this study, the major topics discussed by the Python community are general programming, scientific computing, web development, data structures and formats, installation, deployment, IDE, operating system, multi-threading and multiprocessing, file handling and formats, graph and plotting and GUI support. We notice that the general programming topic has the biggest size, i.e., a large number of questions are asked under this topic. This topic consists of questions like OO principal,

integrating Python with other languages, and implementing simple data structures like lists, dictionaries, etc. The research community should work on developing these concepts further so that it's easier for the software developers to understand. For example, some kinds of transformations or functions can be developed that make it easier to change from one data structure to another.

Universal platform for installation: We notice several questions related to installation issues. Hence, some more research is needed that can make it easier for Python developers to install the packages in an easier manner. From the questions, we notice that there are several ways in which software developers can install packages, for example, using pip, and conda. Also, there are differences in different operating systems like Linux and Windows. Hence, some unified tools are needed that can make package installation easier. Hence, some research is needed to develop a universal package installation system.

OS, multi-threading, and Multiprocessing issues: One of the most problematic subjects for the Python community were OS, multi-threading, and Multiprocessing. A manual examination of the sample questions indicated that several of them had no accepted answer. Although this issue has received considerable attention from the research community [30, 31], it remains a challenging problem in the Python community.

Python multipurpose language: We notice that Python is a multi-purpose language because it is being used in many applications like web development, graph plotting, scientific application, database connections, and as an introductory programming language. Because the software developers are working on so many applications, all these areas have their individual issues as well. Hence, more detailed research is needed that can address the list of issues faced by developers in each of these domains. The are some introductory studies in these areas, for example, analysis of popular topics in Pandas library [22]. More such detailed studies are needed that can help in uncovering specific issues faced in each domain.

4.2. Educator community takeaway

The educational community can leverage the topics obtained from the topic modelling results as a foundation for designing and developing Python courses. The analysis of the topic of general programming revealed that developers find it difficult to integrate various libraries and packages into real-world applications while using general Python concepts. Educators can take this into account and develop courses that focus on **practical applications and demonstrations of appthelying various concepts at an advanced level**. Furthermore, many programmers encounter various errors while coding, therefore **tutorials can be created** with more examples and separate sections for illustrating potential errors, corrective measures, and frequently asked questions in community-based platforms like Stack Overflow. Additionally, more courses on challenging topics like installation, deployments, IDE and OS, multi-threading, and multiprocessing can be developed. More tutorial needs to be developed to understand the different kinds of **data structures** like lists, strings, dictionaries, etc. The **OO principles** also need to discuss in detail.

4.3. Tool/IDE or package development community takeaway

The tool/IDE of the Python package development community plays a vital role in Python software development. Hence there are specific findings that the vendor community can utilize to improve the tools/IDEs/Packages.

Outputting more meaningful or clear error messages: It is noticed from some of the questions that when there is some error in the code it takes lots of time for software developers to figure out the reason for the error because the error message was very generic. Hence, it becomes difficult to debug the code. Providing more clear or more meaningful error messages can help the software development community.
Fast and efficient implementation: We notice several questions where developers were searching for fast or more efficient implementations of data structures or loops. Providing developers with more efficient implementations can be beneficial.
Easier integration with other programming languages: We notice several questions where software developers reported difficulties in the integration of Python with other programming languages or tools like C, C++, SQL, AWS, etc. Hence, the vendor community should work on making these interactions easier and developing APIs that are easier and more robust to use.
Installation of libraries using pip/conda: There are several queries from the software developers that about the installation of libraries using pip or conda. Software developers were facing difficulties in installing packages on different versions of pip. They were also facing difficulty when they had to install multiple packages and they had a dependency on some specific version or system. Hence, the Python development community should focus on making these installation commands more robust or easier so that software developers can easily install the required packages. These installation needs to be tested on various platforms like Windows, Linux, and MacOS Several developers reported issues with package installation in virtual environments or accessing the already installed packages in virtual environments. The tool development community should also work on making the virtual environment creation more robust.
Backward compatibility of the software packages: Software developers report several issues arising because of non-backwards compatibility of functions. They report errors where a previously working code broke because of an update in the packages. Hence, the Python package development community should focus on making the packages backwards compatible.

4.4. Developer community

The findings of this study can be used by Python developers to take a much more disciplined approach to designing diverse Python applications. Organizations can make use of the findings to ensure that the development team receives the proper training on the required Python tools and resources. These findings should be used by the development community to create better tutorials and documentation to reduce any early barriers for new developers. Our investigation revealed that installation, deployment, and IDE were the most difficult topics, and the development community can create manuals and guides to support Python users.

Senior programmers can assist junior programmers more with initial environment setup and package installations. OS, multithreading, and multiprocessing are among the topics with a large number of posts without accepted answers. Software managers can take this into account and allocate more time and resources to finish these tasks.

Additionally, the current documentation for the libraries and modules in these subjects can be improved. On the other side, growing the amount of material can make it more difficult to locate relevant references. Therefore, many of the challenges mentioned in those questions can be addressed by an automated system that recommends appropriate documentation.

5. Threats to validity

In this section, we discuss various threats that can potentially impact the results of our study. We divide this section into three parts: internal validity, external validity and construct validity [32].

5.1. Internal validity

These are validity concerns that may have an impact on our results. The experimental dataset for this study was generated by extracting and analyzing questions from the Stack Overflow website that consists of at least one of the 174 tags that were identified in our "Python-related" tag list (refer to Section 2.2). These tags were determined by manual analysis of questions and after online research into how the tags relate to Python. To further remove the possibility of bias or false positive/false negative in the selection of tags, two of the authors verified the list of tags. Both of these authors had a detailed discussion in case of disagreement. A tag is included in the list only if both of the authors agree to include that tag otherwise the tag is removed from the list. We extracted all the question that consists of at least one of these tags. Using this approach, we extracted 2 449 567 questions, which is comparable to previous approaches that worked on Python questions posts analysis on Stack Overflow [3]. We decided not to include questions that consist of the "Python" in either title or body but do not consist of any of the identified tags from our "Python-related" tag list. Even though this approach will reduce the total number of questions in our experimental dataset, it avoids the possibility of having a large number of false positives as mentioned in the previous approaches [8]. Our approach ensures that we analyze posts that discuss the issues related to Python developers. Additionally, just like in prior studies [8, 20, 33], our analysis is restricted to only looking at the most recent version of the post.

The inclusion of duplicate questions is another validity issue. There are questions on Stack Overflow that seek answers on similar topics or areas, and we haven't used any strategies to separate such questions. The outcomes of this study could be impacted by this. In the future, we intend to broaden our investigation by comparing the outcomes after eliminating duplicates. We also do not analyze the comments associated with Python posts. Comments on questions consist of temporary post-it notes and only privileged users can post comments on a question [8].

In RQ 1.3, we notice that all the questions got answered within 24 hours. We consider a sample of all Pythom questions in this paper. All of those questions in this sample got the answer within 24 hours. However, if we consider the real world there can be a question that took much more time to get the answer. Our dataset included 2,400,000 questions and 26,000,000 answers. To answer the research questions (RQ 1.3, RQ 1.4) we had to run two queries each on all of these questions and one single query was taking around 50 seconds. So, processing the entire set of questions for all the answers was not feasible and would have taken years with our processing capability. So, we had to take a sample set of questions (16 529) to explore these questions with a 99% confidence level that the results given by the query on these 16 529 questions are a representation of the entire population. We took this approach to determine the sample as it was used in many previous papers [12, 13]. We did all the cautions while selecting the sample but still, there can be some bias in the results because of the specific sample that we used. We plan to do more detailed research on a bigger dataset to mitigate this issue.

In addition, in our analysis of popular and difficult questions, we considered questions that are not answered as difficult. In Stack Overflow, questions might be too long or too short, a duplicate of other, ambiguous, and with insufficient details. On Stack Overflow, we frequently see comments asking for more information on questions. These could be some of the reasons it went unanswered or did not have an accepted answer. This might have an impact on the results of our analysis of difficult Python Stack Overflow questions.

In this study, we chose a k value of 20 as the ideal number of topics for topic modelling. The LDA model's ability to provide high-quality topics is directly impacted by this value. It is known that determining an ideal number of "k" is tough. To mitigate this risk, experimentation is carried out with a range of k values ranging from 2 to 100 in increments of two, and topic coherence is calculated for each k value, as in many previous studies [8, 34]. Further, using a regular expression, any code snippets are eliminated from the text before being fed into the LDA algorithm to get better results from the LDA model. This methodology has also been used in numerous earlier research using the Stack Overflow dataset [20].

To reduce any further threats to internal validity, built-in Python modules such as "nltk" and "sklearn" are employed for data processing. Another risk to validity is the manual labelling of the topics provided by the LDA model. The group names produced as part of prior tag analysis RQ 3 (3.2) were used to categorise the topics, and they were not changed after the topic model results are obtained, to eliminate any unconscious biases that might have affected the approach. In addition, a manual examination of the generated topics was performed, and similar topics were merged, when necessary, with relevant labels assigned [13].

While analyzing the results of the LDA algorithm, we notice that some of the posts were incorrectly assigned to a different topic. In our analysis of 376 posts, we found 51 (%13.56) false positive posts, i.e., questions that were suited better for other categories. In these false positives, we notice questions from the topics general programming, web, and GUI which were appearing in other topics.

5.2. External validity

This section includes the threats to the generalisation of the findings obtained from this study. This study gathered data for Python question analysis from Stack Overflow, one of the largest sites on the Stack Exchange network featuring conversations on a wide range of programming topics. However, there are numerous other communities where developers can share their problems and ask questions. Stack Overflow is a popular site featuring a variety of domain expertise. But the study can be further improved by including discussions from other platforms or surveying actual Python developers about the difficulties they face. The study's findings also provide snapshots of data that future research can use to assess how the field has changed and how well Stack Overflow is doing. Finally, the study's findings highlight the issues and difficulties that Python developers faced at the time the study was conducted.

5.3. Construct validity

Construct validity examines how theory and observation relate, or whether the measured values correspond to the actual values. The topics that the LDA algorithm produces may not accurately reflect the posts related to those topics. To mitigate this risk, all keywords linked with each subject were analysed, and a random collection of documents or questions

associated with each topic were verified on the Stack Overflow website before giving a label to each topic to appropriately depict the topic. Additionally, several metrics are employed as part of RQ 4 to assess the popularity and difficulties of key Python topics, and they could pose a risk to the construct validity. However, these measures are employed since they have been widely utilised in many previous studies [8, 13].

6. Related work

In this section, we present an overview of several studies that are closely related to this research. There is a lot of existing work that evaluates various aspects of different programming languages using the Stack Overflow dataset. This section is organised into three main research lines. The studies that examine programming languages using the Stack Overflow dataset are covered in the first category. The second category focuses on studies that analyse the Stack Overflow dataset to identify various software development challenges, and the third category focuses on identifying issues faced by the Python software development community.

6.1. Analysis of Stack Overflow for programming-related questions

In several studies, the Stack Overflow dataset is investigated to examine various aspects of programming languages. The sociological change that the Java developer community has seen is examined in one study which looks at social dynamics with a focus on user altruism and popularity, both internal and external material cross-referencing, and the topics that have generated significant discussions within the Java community over time [11]. The study uses topic discovery to uncover the main discussion topics, and graph mining to depict social interaction within the community.

Another study uses the Stack Overflow dataset and relevant GitHub activity to investigate well-known programming languages like Go, Swift, and Rust [35]. Several research questions are framed to determine the main topics covered, their evolution over time, difficulty level, resource availability, and the relationship between language growth and developer activities.

Other main work uses Python and R to do survival analysis on a Stack Overflow dataset [36]. The emphasis is on predicting the response time for the first answers and accepted answers, where Python had the highest answer rate and R had the highest acceptance answer rate.

Most of the studies mentioned above-analyzed data on Stack Overflow of various programming languages but Python. To the best of our knowledge, only Tahmooresi [3] worked on analyzing Python questions on Stack Overflow. Our study extends and complements their work in several ways. They analyzed topics and tags on Stack Overflow and proposed a tool to detect similar libraries for Python. However, our study examines a variety of factors, including language evolution, themes, user engagement and satisfaction, tags, topics, and topic popularity and difficulty.

6.2. Analysis of Stack Overflow for software development, maintenance, and other aspects

One main study investigates the Stack Overflow dataset to understand the evolution, practices, and challenges faced by programmers in refactoring [8]. The study examines

how refactoring discussions have progressed over time, what are some of the refactoring areas that developers discuss, and which queries are the most popular and challenging. The study looks at the most used tags and terminology in these questions, along with refactor-specific terminologies, as well as the most popular, problematic, and unsolved questions. The goal of the study is to identify areas where refactoring research is lacking, as well as to comprehend why developers' perceptions and actual use of these concepts differ.

Another study used the Stack Overflow dataset to better understand the issues that developers encounter in quantum software engineering (QSE) [12]. They also use GitHub issue reports to learn more about the problems that arise in real-world quantum computing projects. The research focuses on the types of questions about QSE that are posted on the website, QSE topics that are discussed in technical forums, and QSE topics that are highlighted in issue reports of quantum computing projects. The research was able to give insight into future quantum computing potential, uncover several QSE-specific issues experienced by developers, and highlight some of the most popular and difficult quantum computing topics.

An in-depth study uses data from six Q&A Stack Exchange websites to conduct an empirical analysis on logging-related questions [20]. The study looks at the trend of logging questions with accepted answers, the number of answers posted for each question, and the time it takes to receive accepted responses to these questions using statistical analysis. For each of the six websites, the logging questions are also examined in terms of programming languages, with an emphasis on the most common logging questions across programming languages and the distribution of logging questions among programming languages. Content analysis of logging questions is also conducted, with an emphasis on identifying significant topics in the title and description of these questions, as well as analysing the tags associated with these questions. The study shed light on developers' various logging requirements and will aid in the development of better logging tools.

The Stack Overflow dataset is also used in a study on chatbot development issues [13]. This study examines the Stack Overflow posts to learn more about the topics that chatbot developers are interested in and the challenges they face. Topic modelling is used to determine the most popular chatbot-related subjects addressed in the forum and the popularity and difficulty of these topics are also evaluated by the researchers. The report also outlined some conclusions that can be made from it to assist the chatbot community in concentrating on the most pressing issues.

The research discussed above looked into different aspects of software development in various languages. However, none of them concentrated on a comprehensive study of Python language and the various software development issues in the language. Unlike the previous research, this work focuses on analysing the several aspects and challenges of employing Python in software development.

6.3. Analysis of issues faced by Python software development community

The study by Peng [37], performs an empirical investigation on thirty-five prominent Python projects to analyse the use of language features in the projects and develops PYSCAN, an automatic language feature recognizer. The study identifies the top five language characteristics and finds that different domains focus on different language features, with exception handling and nested functions being used most differently. An in-depth examination of these features is performed to summarise their application areas and benefits.

Another study related to Python performs an in-depth analysis of the addition of mutation to Python source code [38]. The mutation method is challenging in dynamically typed languages since the mutant cannot be checked before run time and the applicability of many mutation operators used in other languages is unknown in these dynamically typed languages. The study offers a collection of mutation operators suitable for Python, along with guidelines for choosing operators and a thorough explanation of each. The Python tool for testing mutations, MutPy, is used to implement and analyse each of the discussed operators.

In another major study, a quantitative analysis of Python's performance overheads is carried out with an emphasis on hardware features and language runtime [39]. Additionally, a thorough analysis of how the runtime interacts with the processor's underlying microarchitecture reveal that CPython and PyPy exhibit limited instruction-level parallelism. The paper also discusses several key observations regarding Python's performance optimization possibilities.

All of the above studies concentrate on analysing a single aspect of Python, such as the addition of mutation operations, performance overheads, and language features. The main distinction between our study and others is that our investigation is not constrained to a single language characteristic or paradigm. We attempt to understand the difficulties encountered in the real world by developers while using the language in their applications, as well as some of the areas of Python that are popular and challenging for them.

7. Conclusion

In this empirical study, we conducted a quantitative and qualitative analysis of Python questions posted in Stack Overflow. The quantitative approach involved applying various statistical measures to the extracted data while qualitative analysis involved content analysis of a statistically significant sample of Python questions. Seven research questions were identified, aiming to gain a deeper understanding of the novel challenges encountered by the developers while using Python language in various applications.

The result of this study shows that Stack Overflow is an immensely popular platform where developers seek help and there has been a growing increase in the number of Python questions on the platform from 2008 with a slight drop in 2021. Secondly, analysis of Python tags revealed that the number of questions posted against pandas and TensorFlow is increasing, indicating that scientific computing is becoming a more popular area among Python developers. Thirdly, topic modelling revealed that most of the questions are posted against general programming on concepts like object-oriented principles, algorithm implementation, error management, functional programming, time and date operations, and general Python concepts. Scientific computing, Web development, and Data structures and formats were other topics that developers need assistance with. Finally, the topic of Installation, deployments, and IDE had more view count and was more challenging for developers. Data structures and Formats were identified as the least challenging and most popular topics, with a higher average score and more satisfactory response from the community. This study opens doors for Python researchers and practitioners to further comprehend Python development challenges. From the results obtained from this study, a series of actionable takeaways are highlighted for relevant stakeholders that can improve the field of Python programming and enhance developer productivity.

In the future, we plan to extend topic modelling on the entire dataset to perform a comparison with the results of this study and to analyze how Python topics evolve. At

present, we analyzed popular topics based on three metrics, i.e., average view count, average score, and average favourite count. In the future, we plan to consider the size of a topic (#number of questions) as one of the popularity metrics. Also, a structured survey with junior and senior Python developers can be conducted, exploring their general and specific challenges encountered while using Python. This survey can complement and evaluate this Stack Overflow study to provide the software engineering community with a more comprehensive view of Python development practices and difficulties. Finally, the study can be expanded to investigate developers' discussions on other forums and repositories to evaluate the commits and defect reports to obtain further insights regarding numerous problems faced by Python developers.

Funding

This work is not supported by any external funding.

Acknowledgement

We thank Harshit Gujral for helping us with the LDA code.

References

[1] M. Lutz, *Programming Python*. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly Media, Inc., 2001.

[2] K. Chowdhary, "On the evolution of programming languages," *arXiv preprint arXiv:2007.02699*, 2020.

[3] H. Tahmooresi, A. Heydarnoori, and A. Aghamohammadi, "An analysis of Python's topics, trends, and technologies through mining Stack Overflow discussions," *arXiv preprint arXiv:2004.06280*, 2020.

[4] B.A. Malloy and J.F. Power, "An empirical analysis of the transition from Python 2 to Python 3," *Empirical Software Engineering*, Vol. 24, No. 2, 2019, pp. 751–778.

[5] Z. Zhang, H. Zhu, M. Wen, Y. Tao, Y. Liu et al., "How do Python framework APIs evolve? An exploratory study," in *27th international conference on software analysis, evolution and reengineering (saner)*. IEEE, 2020, pp. 81–92.

[6] R. Widyasari, S.Q. Sim, C. Lok, H. Qi, J. Phan et al., "BugsInPy: A database of existing bugs in Python programs to enable controlled testing and debugging studies," in *Proceedings of the 28th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 2020, pp. 1556–1560.

[7] A. Tashakkori, C. Teddlie, and C.B. Teddlie, *Mixed methodology: Combining qualitative and quantitative approaches*, Vol. 46, Thousand Oaks, California: Sage, 1998.

[8] A. Peruma, S. Simmons, E.A. AlOmar, C.D. Newman, M.W. Mkaouer et al., "How do I refactor this? An empirical study on refactoring trends and topics in Stack Overflow," *Empirical Software Engineering*, Vol. 27, No. 1, 2022, pp. 1–43.

[9] K. Georgiou, N. Mittas, A. Chatzigeorgiou, and L. Angelis, "An empirical study of COVID-19 related posts on Stack Overflow: Topics and technologies," *Journal of Systems and Software*, Vol. 182, 2021, p. 111089.

[10] G. Pinto, F. Castor, and Y.D. Liu, "Mining questions about software energy consumption," in *Proceedings of the 11th Working Conference on Mining Software Repositories*, 2014, pp. 22–31.

[11] G. Blanco, R. Pérez-López, F. Fdez-Riverola, and A.M.G. Lourenço, "Understanding the social evolution of the Java community in Stack Overflow: A 10-year study of developer interactions," *Future Generation Computer Systems*, Vol. 105, 2020, pp. 446–454.

[12] H. Li, F. Khomh, M. Openja et al., "Understanding quantum software engineering challenges an empirical study on Stack Exchange forums and GitHub issues," in *International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2021, pp. 343–354.

[13] A. Abdellatif, D. Costa, K. Badran, R. Abdalkareem, and E. Shihab, "Challenges in chatbot development: A study of Stack Overflow posts," in *Proceedings of the 17th international conference on mining software repositories*, 2020, pp. 174–185.

[14] W. McKinney et al., "pandas: A foundational Python library for data analysis and statistics," *Python for High Performance and Scientific Computing*, Vol. 14, No. 9, 2011, pp. 1–9.

[15] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digital Journalism*, Vol. 4, No. 1, 2016, pp. 89–106.

[16] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," *To Appear*, Vol. 7, No. 1, 2017, pp. 411–420.

[17] Y. Zhang, R. Jin, and Z.H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, Vol. 1, No. 1, 2010, pp. 43–52.

[18] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, No. 4/5, 2003, pp. 993–1022.

[19] R.H. Ali and E. Linstead, "Modeling topic exhaustion for programming languages on Stack Overflow," in *SEKE*, 2020, pp. 400–405.

[20] H. Gujral, A. Sharma, S. Lal, and L. Kumar, "A three dimensional empirical study of logging questions from six popular Q & A websites," *e-Informatica Software Engineering Journal*, Vol. 13, No. 1, 2019.

[21] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[22] S.K.S. Joy, F. Ahmed, A.H. Mahamud, and N.C. Mandal, "An empirical studies on how the developers discussed about pandas topics," *arXiv preprint arXiv:2210.03519*, 2022.

[23] J.C. Westland, "The cost of errors in software development: Evidence from industry," *Journal of Systems and Software*, Vol. 62, No. 1, 2002, pp. 1–9.

[24] S. Ahmed and M. Bagherzadeh, "What do concurrency developers ask about? A large-scale study using Stack Overflow," in *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*, 2018, pp. 1–10.

[25] K. Bajaj, K. Pattabiraman, and A. Mesbah, "Mining questions asked by web developers," in *Proceedings of the 11th Working conference on mining software repositories*, 2014, pp. 112–121.

[26] M. Bagherzadeh and R. Khatchadourian, "Going big: A large-scale study on what big data developers ask," in *Proceedings of the 2019 27th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 2019, pp. 432–442.

[27] S. Nadi, S. Krüger, M. Mezini, and E. Bodden, "Jumping through hoops: Why do Java developers struggle with cryptography APIs?" in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 935–946.

[28] X.L. Yang, D. Lo, X. Xia, Z.Y. Wan, and J.L. Sun, "What security questions do developers ask? A large-scale study of Stack Overflow posts," *Journal of Computer Science and Technology*, Vol. 31, No. 5, 2016, pp. 910–924.

[29] C. Rosen and E. Shihab, "What are mobile developers asking about? A large scale study using Stack Overflow," *Empirical Software Engineering*, Vol. 21, No. 3, 2016, pp. 1192–1223.

[30] A. Malakhov, D. Liu, A. Gorshkov, and T. Wilmarth, "Composable multi-threading and multi-processing for numeric libraries," in *Proceedings of the 17th Python in Science Conference, Austin, TX, USA*, 2018, pp. 9–15.

[31] Q. Nguyen, *Mastering Concurrency in Python: Create faster programs using concurrency, asynchronous, multithreading, and parallel programming*. Birningham, UK: Packt Publishing, Ltd., 2018.

[32] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell et al., *Experimentation in software engineering*. Springer Science and Business Media, 2012.

[33] H. Gujral, S. Lal, and H. Li, "An exploratory semantic analysis of logging questions," *Journal of Software: Evolution and Process*, Vol. 33, No. 7, 2021, p. e2361.

[34] M. Alshangiti, H. Sapkota, P.K. Murukannaiah, X. Liu, and Q. Yu, "Why is developing machine learning applications challenging? A study on Stack Overflow posts," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 2019, pp. 1–11.

[35] P. Chakraborty, R. Shahriyar, A. Iqbal, and G. Uddin, "How do developers discuss and support new programming languages in technical Q&A site? An empirical study of Go, Swift, and Rust in Stack Overflow," *Information and Software Technology*, Vol. 137, 2021, p. 106603.

[36] L. Lord, J. Sell, F. Bagirov, and M. Newman, "Survival analysis within Stack Overflow: Python and R," in *4th International Conference on Big Data Innovations and Applications (Innovate-Data)*. IEEE Computer Society, 2018, pp. 51–59.

[37] Y. Peng, Y. Zhang, and M. Hu, "An empirical study for common language features used in Python projects," in *International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2021, pp. 24–35.

[38] A. Derezińska and K. Hałas, "Analysis of mutation operators for the Python language," in *Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX. June 30–July 4, 2014, Brunów, Poland*. Springer, 2014, pp. 155–164.

[39] M. Ismail and G.E. Suh, "Quantitative overhead analysis for Python," in *International Symposium on Workload Characterization (IISWC)*. IEEE, 2018, pp. 36–47.

e-Informatica Software Engineering Journal (EISEJ) is an international, fully open access (CC-BY 4.0 without any fees for both authors and readers), blind peer-reviewed computer science journal using a fast, continuous publishing model (papers are edited, assigned to volume, receive DOI & page numbers, and are published immediately after acceptance without waiting months in a queue to be assigned for a specific volume/issue) without paper length limit that concerns theoretical and practical issues pertaining development of software systems. Our aim is to focus on empirical software engineering, as well as data science in software engineering.

The journal is published by *Wrocław University of Science and Technology* under the auspices of the *Software Engineering Section of the Committee on Informatics of the Polish Academy of Sciences*.

Aims and Scope

The purpose of **e-Informatica Software Engineering Journal** is to publish original and significant results in all areas of software engineering research.

The scope of **e-Informatica Software Engineering Journal** includes methodologies, practices, architectures, technologies and tools used in processes along the software development lifecycle, but particular stress is laid on empirical evaluation using well-chosen statistical and data science methods.

e-Informatica Software Engineering Journal is published online and in hard copy form. The on-line version is from the beginning published as a gratis, no authorship fees, open-access journal, which means it is available at no charge to the public. The printed version of the journal is the primary (reference) one.

Topics of interest

- Software requirements engineering and modeling
- Software architectures and design
- Software components and reuse
- Software testing, analysis and verification
- Agile software development methodologies and practices
- Model driven development
- Software quality
- Software measurement and metrics
- Reverse engineering and software maintenance
- Empirical and experimental studies in software engineering (incl. replications)
- Evidence-based software engineering
- Systematic reviews and mapping studies (see SEGRESS guidelines)
- Statistical analyses and meta-analyses of experiments
- Robust statistical methods
- Reproducible research in software engineering
- Object-oriented software development
- Aspect-oriented software development
- Software tools, containers, frameworks and development environments
- Formal methods in software engineering.
- Internet software systems development
- Dependability of software systems
- Human-computer interaction
- AI and knowledge based software engineering
- Data science in software engineering
- Prediction models in software engineering
- Mining software repositories
- Search-based software engineering
- Multiobjective evolutionary algorithms
- Tools for software researchers or practitioners
- Project management
- Software products and process improvement and measurement programs
- Process maturity models

Funding acknowledgements: Authors are requested to identify who provided financial support for the conduct of the research and/or preparation of the article and to briefly describe the role of the sponsor(s), if any, in study design; in the collection, analysis and interpretation of data; in the writing of the paper. If the funding source(s) had no such involvement then this should be stated as well.

The submissions will be accepted for publication on the base of positive reviews done by international Editorial Board and external reviewers.

English is the only accepted publication language. To submit an article please enter our online paper submission site.

Subsequent issues of the journal will appear continuously according to the reviewed and accepted submissions.

The journal is included in the IC Journal Master List (ICV=7.59 was obtained in 2013) and indexed by Scopus, DBLP, DOAJ, BazTech etc.

Paper copies of selected issues of the journal are available from our Publisher (please contact oficwyd@pwr.wroc.pl for details).

http://www.e-informatyka.pl/

e-Informatica

ISSN 1897-7979