Current Practices in Quality Assessment of Systematic Reviews in Computing: Exploring Automation Potential

Ghader Reda Kurdi
Department of Data Science, College of Computing, Umm Al-Qura University, Saudi Arabia
grkurdi@uqu.edu.sa (corresponding author)

Budoor Ahmad Allehyani
Department of Software Engineering, College of Computing, Umm Al-Qura University, Saudi Arabia
balehyani@uqu.edu.sa

Received: 19 November 2024 | Revised: 17 January 2025 | Accepted: 5 February 2025

ABSTRACT

Systematic Literature Reviews (SLRs) play a crucial role in evidence synthesis within computing research. However, the quality of SLRs can vary significantly, affecting their reproducibility and trustworthiness. This study addresses the problem of poorly understood practices in SLR quality assessment. It investigates the current landscape of quality instruments used to assess SLRs in computing by analyzing 97 tertiary studies across various computing domains. The analysis focuses on identifying the dominant quality instruments, and examining reported modifications or adaptations made to them. A qualitative analysis is conducted on the interpretations and scoring of widely utilized quality criteria. The analysis reveals diverse interpretations and potential inconsistencies in the application of quality instruments, owing to the absence of concrete examples. The findings provide valuable insights for both SLR authors and consumers in computing research, pointing out the most widely deployed quality instruments, common customization and interpretive practices, and potential areas for improvement. This study contributes to the ongoing discussions on enhancing SLR quality in computing, forming the basis for automating the quality assessment process.

Keywords-mapping study; quality assessment; quality criteria; reporting quality; reproducibility; systematic literature review; scoping review; tertiary study

I. INTRODUCTION

SLRs aim to identify, evaluate, and consolidate existing literature using a systematic, transparent, and reproducible methodology. Originating in medicine, this review methodology has expanded its influence into various other fields. Since its introduction into software engineering in 2004 [1], the methodology has gained widespread recognition and increased adoption, not only in software engineering, but also in other areas within computing. The SLR process commences with the identification of the question(s) that will guide the review, followed by the formulation of a plan for collecting primary studies, assessing their quality, utilizing them for data extraction, and analyzing and synthesizing them. Once the plan is executed, comprehensive details regarding the execution of each phase and the results should be reported. This emphasis on reporting is a key distinguishing characteristic that sets SLRs apart from other types of literature reviews. Numerous guidelines and checklists are available to assist reviewers in ensuring SLR quality and comprehensive reporting. For example, the Database of Attributes of Reviews of Effects (DARE) utilizes a set of five quality criteria (https://www.crd.york.ac.uk/), and SLRs must meet at least four of them to qualify for inclusion in the database. Another popular instrument is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, which features both a checklist and a flow diagram (https://www.prisma-statement.org/). The checklist comprises 27 items and reports in different sections of the review documents, while the flow diagram mainly focuses on reporting the study identification procedure. Additionally, the National Institute of Health (NIH) provides a dedicated quality assessment tool for SLRs and meta-analyses (https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools), offering another valuable resource for reviewers. Efforts have been made to tailor existing instruments for SLRs to the field of computing, with particular attention being paid to software engineering. Examples of such initiatives include QAISER, which aims to assist appraisers in assessing SLR quality using 16 items [2], and SEGRESS [3], which is more oriented toward qualitative SLRs and mapping studies than the PRISMA 2020. The latter builds upon and was originally designed for quantitative SLRs and meta-analyses of formal experiments and quasi-experiments. However, while guidelines and checklists are available, evaluating the actual SLR quality remains crucial. Different SLR quality aspects, including reporting [6], search strategies [7], threats to validity [8, 9], and contributing factors [10, 11] have been investigated in the existing literature. Authors in [6] examined 37 software engineering SLRs published between 2010 and 2015, identifying reporting quality issues and creating a 9-item checklist to improve essential reporting. Authors in [7] developed a checklist to assess the reliability of automated search strategies in SLRs, finding that the 92% of 27 reviews were not repeatable due to missing search details. Authors in [8] identified common threats to the validity of 316 SLRs and mapping studies, while authors in [9] highlighted two primary threats to replicating SLRs in computer science and offered recommendations for mitigating them. Authors in [10, 11] examined the impact of several factors, such as referencing guidelines, publication venue, and research scope, on SLR quality and concluded that improvements have been made in the latter over time.

Similar to the current study's focus, authors in [12] investigated the assessment of secondary studies' quality conducted as part of tertiary studies in the field of software engineering. Three quality assessment aspects were examined: 1) frameworks in use, 2) the facets explored within these frameworks, and 3) the purposes. To accomplish this, the authors performed a content analysis of 47 tertiary studies obtained through a systematic mapping study conducted on Scopus. They identified the DARE framework as the most widely used quality assessment instrument, with some studies having customized this framework to suit their specific needs. However, a notable discrepancy arises from the lack of comprehensive analysis and synthesis regarding how quality assessment instruments are used and customized within computing. Existing studies focus on specific aspects or smaller samples, generating gaps in understanding how these instruments are applied across the broader computing field. For example, authors in [12] concentrated on SLRs within software engineering obtained from a single source, limiting its scope and generalizability. Moreover, previous studies have not examined how these quality criteria are interpreted and applied across the computing field. The DARE framework, for instance, has been widely used to assess review quality, but prior research has not fully explored whether the quality assessment criteria are being consistently interpreted and applied.

Given that conducting SLRs is a labor-intensive and time-consuming process, there has been a growing interest in automating the procedure. Several studies have explored automating various tasks involved in SLRs, with the greatest attention having been paid to screening (i.e. the selection of relevant studies) [4]. However, one aspect that has not received sufficient attention is automating the evaluation of the reporting quality in SLRs themselves, a crucial factor for enhancing the reproducibility and trustworthiness of their findings. This is particularly significant, given that various studies have pinpointed limitations in the quality of different SLR aspects, such as search strategy, synthesis, and reporting [5].

Given the increasing reliance on SLRs across various disciplines within computing, ensuring that these reviews maintain high quality is crucial for advancing the field. Automating the assessment of the SLR reporting quality may enable quicker evaluations and reduce human bias or inconsistencies. Motivated by the automation potential, this research aims to gain a deeper understanding of the assessment practices used to evaluate published SLRs and to reflect on possible ways for these practices to be automated. Consequently, the following research questions were formulated:

RQ1: What are the existing instruments for rating the SLR quality in computing?

RQ2: Which instruments are most widely used?

RQ3: Are the quality criteria within these instruments employed "as-is" or customized to suit computing?

RQ4: How have researchers interpreted the quality criteria within the applied instruments?

The main contribution of this research is to provide a comprehensive 14-year overview of the SLR quality assessment. It has identified the most frequently used instruments and examined the aspects adapted to align with the specific needs of the computing field. One of the key results of this study is highlighting inconsistencies in the interpretation of DARE quality criteria. In response, it has developed a set of recommendations aimed at addressing these inconsistencies and streamlining the quality assessment process, laying the groundwork for automating the latter. The information presented in this paper has significant value for researchers interested in implementing quality assessments to evaluate SLRs and those exploring the possibility of automating the SLR quality assessment process.

II. RESEARCH METHODOLOGY

A. Data Source

In this work, the primary source of information on quality was derived from tertiary studies, which are SLRs of secondary studies; that is, SLRs and mapping studies. These tertiary studies were obtained from the resource for computing-related systematic secondary studies [13]. The resource currently includes 4,217 systematic secondary studies, comprising SLRs, mapping studies, and tertiary studies.

B. Procedure

To specifically identify tertiary studies, R was utilized to filter in studies explicitly containing the keyword "tertiary" in their title, abstract, or keywords. The results were exported into a CSV file and the resulting 99 tertiary studies were reviewed to confirm their classification as tertiary studies. Following this initial review, one study was omitted due to the inaccessibility of the full text, and another one was excluded because it was not in English. This resulted in a dataset comprising 97 tertiary studies for subsequent data extraction and analysis. The following data about the SLR quality assessment were extracted.

- Basic data about tertiary studies

  - Type of the tertiary study
  - Computing area
  - Publication year
  - Number of reviews included in the tertiary study
- Data about quality instruments
  - Instruments used for quality assessment
  - Number of quality criteria
  - Modification of the quality assessment instrument, including the incorporation of additional quality criteria
  - Interpretation of the quality criteria

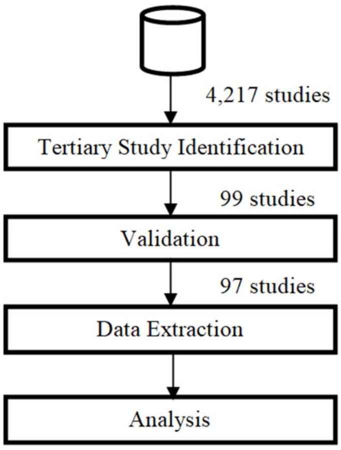

Fig. 1. Study workflow.

Regarding the instruments used for quality assessment, some research indicates the utilization of DARE, while other studies point to adopting Kitchenham evaluation criteria. However, the criteria outlined in Kitchenham were originally derived from DARE, albeit with refinements made through adjustments in phrasing. Table I illustrates the differences between the evaluation criteria of Kitchenham and DARE. Notably, some studies use the phrasing of Kitchenham but refer to DARE. To maintain consistency and accuracy in data extraction, this study adheres to referencing DARE in the extraction table, while acknowledging that the wording of Kitchenham's criteria is utilized. The present work's assessment was based on the quality criteria reported in the reviewed studies and not on the references provided by the researchers.

Data were extracted in tabular formats and analyzed both manually and employing R. For example, R was used to preprocess keywords associated with tertiary studies and generate a word cloud, providing insights into the most prevalent topics and computing areas. The preprocessing involved standardizing the text by converting it to lowercase, eliminating common stopwords, and removing keywords linked to SLRs, such as "tertiary studies", "systematic literature review", and "systematic mapping". This step was taken to ensure that the generated word cloud distinctly reflects the computing topic explored within the studies.

TABLE I. DARE AND CORRESPONDING KITCHENHAM EVALUATION CRITERIA

DARE	Kitchenham et al. [14]
Were inclusion/exclusion criteria reported?	Are the review's inclusion and exclusion criteria described and appropriate?
Was the search adequate?	Is the literature search likely to have covered all relevant studies?
Was the quality of the included studies assessed?	Did the reviewers assess the quality/validity of the included studies?
Are there sufficient details about the included studies presented?	Were the basic data/studies adequately described?
Were the included studies synthesized?	

Additionally, R was utilized to analyze basic data about the tertiary studies. That is, it was used to obtain summary information about the publication years, as well as the number of reviews included in tertiary studies, such as the minimum, maximum, and total number across all studies. R was also utilized to count the number of occurrences of different quality instruments. On the other hand, certain aspects, involving the modification of quality instruments and the interpretation of quality criteria, required manual analysis. For instance, researchers might use different descriptions but convey the same meaning when assigning certain scores during quality assessment. There are also instances where very similar descriptions are used with slight differences that alter the meaning. To address this, a thorough manual analysis was conducted by reading through these interpretations, grouping similar ones, and then counting to determine the prevalence in the interpretations. Similarly, this study delved into the nature of modifications made to quality assessment instruments, examined them, and extracted common themes deploying the coding framework outlined in [12].

III. RESULTS

A. *Basic Information*

Initially, basic information about the dataset was provided. A total of 97 tertiary studies published between 2009 and 2023 (listed in Table II) were analyzed, covering a cumulative 2,620 secondary studies. This count excludes the study reported in [15] due to its ongoing status. Among tertiary studies, 22 were identified as SLRs, three as systematic mappings, and one as an interpretative review.

Software engineering emerged as the most prominently represented sub-area. This observation is further emphasized through the word cloud generated from the keywords, as illustrated in Figure 1. Within this visual representation, the terms "software" and "engineering" stand out as the most frequently occurring. Additionally, words like "agile" and "development", often associated with concepts, such as "Agile Software Development", "Distributed Software Development", or "Global Software Development", contribute to affirming the substantial focus on this specific domain.

Fig. 2. Word cloud of keywords extracted from the tertiary studies.

B. Current Instruments

Quality assessment was conducted in 50 studies. Regarding the quality assessment instruments, DARE was utilized in 84% (n=42) of the studies performing quality assessment. It is noteworthy that there are both old (DARE-4) and new versions of DARE (DARE-5), with the key distinction being that the new version incorporates a fifth quality criterion added in 2009, [16]. The most commonly employed version was DARE-4 (25 studies) followed by DARE-5 (17 studies). All studies in the performed analysis, except for [14], were published post-2009, after DARE-5 became available. However, DARE-4 is still more frequently used, highlighted by the fact that 30 studies adopted the language utilized in [14], which deployed DARE-4. This may explain the popularity of DARE-4 over DARE-5. It should be mentioned, though, that of the 30 studies that adopted the language used in [14], 22 utilized DARE-4, while 8 employed DARE-5. The remaining 16% (n=8) of the studies utilized the well-known PRISMA (1 study), AMSTAR (1 study), and custom-defined checklists (6 studies) for quality assessment (Table III), with some referring to other studies for guidance in formulating these checklists (i.e. [17] referring to [6, 10, 18] and [19] referring to [20, 21]).

C. Modifications of Quality Instruments

It is important to note that Kitchenham et al. [14] have modified the wording of the DARE criteria and provided interpretations of the DARE criteria to facilitate the assignment of scores for each criterion. The original DARE criteria and the adapted versions are presented in Table I. Interestingly, 73.17% (n=30) of the studies utilizing DARE employ the wording introduced in [14]. Some studies refer to [14] without explicitly citing DARE, thus supporting this work's speculation regarding its popularity. Apart from modifications to the wording, the DARE-4 was used and underwent modifications in only five tertiary studies, whereas DARE-5 was modified in two [26, 31]. In all studies that involved instrument modification, a consistent pattern emerged wherein new criteria were introduced. However, in [27], the criterion "is the literature search likely to have covered all relevant studies?" was omitted from DARE, while additional criteria were incorporated. It was found that items added repeatedly pertained to research questions (QC6), study types (QC22), potential threats to validity (QC8), precise reporting of the findings (QC28), and precise reporting of research method (QC29).

D. Interpretation of Quality Instruments

1) Interpretation of the Quality Criteria related to the Inclusion/Exclusion Criteria

Concerning the interpretation of the first DARE criteria, specifically related to the inclusion/exclusion criteria, the analysis revealed that in [10, 11, 14, 16, 26, 28, 32-46], the focus was primarily on the inclusion criteria alone. If the inclusion criteria are explicitly stated, the paper is assigned a full score; if implicit or partially described, it receives a partial score, and if not defined, a score of 0 is assigned. Furthermore, in [46], it is asserted that the explicit inclusion criteria are those clearly delineated in a separate section of the paper.

In [47-49], it is argued that the complete satisfaction of the first DARE criteria requires an explicit statement of either inclusion or exclusion. In contrast, implicit representation results in partial satisfaction [6, 48-51], and the absence of a definition leads to non-satisfaction [48, 49]. In [51-55], both criteria need to be explicitly addressed to meet the specified DARE criteria. Within these studies, there is a variation in the interpretation of a partial score. Different interpretations include scenarios where the criteria are implicit but safely inferred, either criterion is implicit, both are implicit, both are partially defined, only one selection criterion is described, or they are both implicit. In [6, 30, 50, 56, 57], references to the criteria were made without inclusion or exclusion criteria having been explicitly specified. The condition for assigning a full score includes explicit criteria, while for a partial score, involves criteria that are implicit or partially defined. Authors in [30] provide an additional elaboration on the notion of "explicit criteria", stating that they are presented in tables, bullet points, or clearly described in the text, while implicit criteria are derived from research questions and search terms without explicit clarification and a zero score is assigned if the criteria are not defined.

While many studies adopted the version "are the review's inclusion and exclusion criteria described and appropriate?", found in [14], instead of the original DARE criteria "were inclusion/exclusion criteria reported?", it is evident from their interpretation that they all focus exclusively on reporting the criteria without assessing their appropriateness. Establishing universal rules for assessing appropriateness is challenging, as it depends on the topic and objectives of the reviews. However, authors can provide interpretations of appropriateness that are tailored to the specific context of their review.

In addition, most studies concentrated primarily on inclusion criteria. This exclusive focus can lead to a limited understanding of the selection process, as exclusion criteria are equally important to ensure the rigor and reproducibility of an SLR.

TABLE II. BASIC INFORMATION ABOUT INCLUDED STUDIES. SE = SOFTWARE ENGINEERING, TS = TERTIARY STUDY, SM = SYSTEMAIC MAPPING

Reference	Domain and focus	Type of review	No. of included reviews
[14]	SE (SLRs in SE)	SLR	20
[65]	SE (Synthesis in SLRs in SE)	TS	31
[11]	SE (SLRs in SE)	SLR	67 but 42 selected for quality assessment
[10]	SE (SLRs in SE)	SLR	33
[66]	Health information systems	SLR	50
[67]	SE (Research synthesis in SE)	SLR	49
[33]	SE (SLRs in SE)	SLR	77
[59]	SE (Evidence-informed teaching in SE)	SLR	48
[68]	SE (Communication in distributed development)	SLR	20
[34]	SE (Distributed software development)	SLR	14
[35]	SE (GSD)	SLR	37
[16]	SE (GSD)	TS	37
[69]	SE (Agile and lean practices in SE)	TS	13
[8]	SE (Validity of SLRs)	TS	316
[36]	SE (Software product lines)	TS	60
[70]	Automation in code generation	TS and SM	2,450
[6]	Reporting SLRs in SE	TS	37
[27]	Internet of things	TS	12
[71]	SE (Software productivity)	TS	4
[72]	SE (Threats to validity in SE secondary studies)	TS	165
[28]	SE (Software reuse)	TS	56
[73]	SE (Software visualization)	SLR	48
[74]	Cloud computing	SLR	76
[52]	SE (Gamification)	SM	12
[75]	SE (Multivocal literature reviews)	TS	12
[76]	SE (Model-driven engineering)	SLR	64
[77]	SE (GSD)	TS	25
[17]	AI Adoption in business and management	SLR	45
[56]	SE (Requirement patterns)	SLR	4
[78]	SE (Education)	TS	26
[37]	Sentiment analysis	SLR	16
[57]	SE (software testing)	SLR	49
[24]	SE (Code smells)	TS	13
[25]	SE (Quality assessment in SLRs)	TS	241
[79]	SE (DevOps)	SM	41
[38]	Cyber bullying-Cyber harassment	SLR	50
[39]	SE (Software process improvement)	TS	24
[58]	SE (Scaling agile development process)	TS	7
[29]	Internet of things	TS	11
[47]	SE (Requirements engineering)	TS	53
[80]	Mixed integer programming	Literature review and TS	18
[50]	SE (teaching)	TS	49
[81]	SE (Using grey literature)	TS	9
[40]	SE (Usability in agile software development)	TS	14
[82]	SE (Evidence-based software engineering research)	TS	19
[15]	SE (Grey literature reviews)	TS	13,177
[83]	SE (Agile practices)	TS	37
[84]	Fourth industrial revolution of supply chains	TS	65
[41]	Cybersecurity behavioral research	TS	107
[85]	SE (SLRs)	TS	48
[42]	SE (Agile software development)	TS	28
[19]	SE (Test case selection and prioritization)	TS	22
[86]	SE (Meta-ethnographies)	TS	44
[22]	SE (GSD)	TS	27
[60]	SE (Software testing)	TS	53
[87]	Sentiment analysis	SLR	22
[88]	SE (Pareto's Law)	TS	107
[89]	SE (Grey literature)	TS	446
[90]	SE (Searches in secondary studies)	TS	50
[91]	Supply chain management	SLR	74
[30]	Software product lines and variability modeling	TS	86
[26]	Software ecosystems	TS	22
[92]	SE (Agile trends in SE)	TS	12

[9]	Threats to replicating SLR searches	SLR	289
[93]	Computer science technologies for autism	SM	33
[23]	SE (Experiences of conducting SLRs in SE)	TS	116
[43, 44]	SE (Software testing)	TS	22
[45]	Technical debt	TS	13
[94]	SE (Agile software development)	TS	118
[51]	System of systems Architecture	TS	19
[95]	Blockchain and sustainability	TS	34
[48]	Technical debt management	TS	19
[31]	SE (Software cost/effort estimation)	TS	14
[49]	Machine learning for SE	TS	83
[96]	Machine scheduling problems in production	TS	129
[32]	SE (Software process improvement)	TS	70
[97]	Industry 4.0	TS	46
[98]	SE practice	TS	120
[99]	No/Low-added value technology	Interpretative review	17
[100]	Data analytics in healthcare	TS	45
[46]	Technology use in education	TS	73
[101]	Technology use in classrooms	Tertiary meta-analytic review	NA
[55]	SE (Software development teams)	TS	29
[102]	Types of interoperability	TS	15
[53]	Systematic mapping in SE	SLR	178
[103]	Microservices anti-patterns	TS	7
[104]	Microservice architecture	TS	37
[105]	SE (Requirements in engineering practices)	TS	5
[106]	Grey literature and google scholar in SLR in SE	TS	138
[107]	SLRs and systematic mapping in SE	TS	170
[108]	Software quality measurement	TS	75
[109]	Quality assessment in SE	TS	127
[54]	Model-driven engineering	TS	22
[110]	Blockchain	TS	42
[111]	SE (Code smells and refactoring)	TS	21
[112]	SE (variability modeling)	TS	78

2) Interpretation of the Quality Criteria related to the Adequacy of the Search

Regarding the adequacy of the search in [6, 10, 14, 16, 26, 28, 30, 32-35, 37, 40, 45, 48, 50, 57], two conditions emerged as indicators of search adequacy: 1) employing two search strategies, with one being searches in at least four electronic databases or 2) identifying and searching all relevant journals on the subject of interest. In [43, 44], the conditions for achieving a full score are similar, although for the first condition, the authors specify that the number of digital libraries is one.

Even though many studies assert that searching relevant journals alone is deemed sufficient without incorporating the first condition, in [36], a broader approach is followed. This study specifies that adequacy requires the utilization of two complete search strategies. It offers an illustrative example of complete strategies, emphasizing the need to search at least four digital libraries and conduct a manual search across all potential forums. Notably, authors in [39, 46, 47, 51, 52, 54, 56] specify only the first condition, while authors in [49, 55] adopt a similar condition but specify that the number of databases should be more than four.

Furthermore, in [38, 41, 42, 53], searching databases alone is deemed sufficient. These studies explicitly specify that to achieve a full score, the searches must encompass four or more academic and reputable online databases. For a partial score, three to four databases are required [41, 42, 53], or fewer than four [38]. A score of 0 is given if two or fewer databases are searched.

In [6, 10, 14, 16, 26, 32, 34, 35, 37, 40, 45, 48, 50, 57], a partial score is assigned when three or four databases are searched without applying additional search strategies or when the search is limited to a restricted set of journals. The remaining studies exhibit varied overlapping criteria for assigning partial scores. In [28], a partial score requires searching across two or three digital libraries or searching a defined but restricted set of journals and conferences. In [30], the approach to assigning partial scores differs, specifying the condition as searching four digital libraries without employing additional search strategies or searching three digital libraries with the incorporation of extra search strategies. Authors in [33] adopted similar criteria for a partial score to those presented in [30], with a slight variation in searching three databases regardless of the use of additional strategies. They also introduce an additional criterion involving a search within a restricted set of journals and conference proceedings. Finally, authors in [36] conditioned a complete primary search strategy, an incomplete one, or no secondary search strategy.

For a partial score, some studies require searching fewer databases than what is necessary for a full score. For instance, searching three to four databases is deemed sufficient in [39, 47, 49, 51, 52, 54, 58], while searching three relevant sources is adequate in [56]. A score of zero is assigned for searches conducted with up to two digital libraries or an extremely restricted set of journals [6, 10, 14, 16, 26, 28, 30, 32- 37, 40, 45, 48, 50, 52, 57]. In [58], the criterion is similar but includes searches in two databases. Authors in [10, 37] emphasize the importance of evaluating the appropriateness of digital libraries for a specific SLR, although they do not specify the exact methodology used for this evaluation.

3) Interpretation of the Quality Criteria related to the Synthesis Method

For synthesis, a full score is assigned in [16, 26, 35, 40, 48, 53] when "an explicit synthesis method is named, and a reference to the method is supplied". A partial score is given if the synthesis method is named, but no reference to it is supplied. Authors in [41] claim that in order to assign a full score, "the synthesis method is explicitly defined". However, based on the condition for assigning a partial score, according to which, "the synthesis method is implicitly mentioned without any reference", it appears that the authors apply the same criteria as the six aforementioned studies. Nevertheless, it remains unclear which reference should be provided, as commonly known synthesis methods, such as meta-analysis, typically do not necessitate a specific citation.

Authors in [50, 59], assign a full score for conducting synthesis or meta-analysis for all the data and a partial score when synthesis or meta-analysis are performed for only some of the data. However, synthesizing all of the data is sometimes impractical due to their heterogeneity, a limitation which contributes to improving the synthesis quality. Authors in [55] appear to employ the same criteria, assigning a full score when the results are compared with those of other studies and a partial score when only a few results are discussed. In [57], two sub-questions were derived from the DARE criteria to assess the synthesis appropriateness. The first sub-question addresses whether the data are both synthesized and aggregated or merely summarized. The second sub-question focuses on the quality of the studies included in the synthesis. Similarly, the interpretations employed in [51] appear to align with the first sub-question, awarding a full score if the data are extracted, summarized, and interpreted. A partial score is awarded if the data are not interpreted, and no score is given if the data are not summarized.

4) Interpretation of the Quality Criteria related to the Quality/Validity

When evaluating quality assessment in [10, 14, 16, 26, 28, 32-37, 39, 40, 43-47, 49-54, 56, 57, 59], the predominant criterion for assigning full scores involves the explicit definition and extraction of quality criteria from each reviewed study. In [26], the same criterion is applied, but the incorporation of preventative steps is also required to minimize bias and errors in the quality assessment process. It is worth mentioning that in [39, 47, 49], the phrasing varies, with two of them using "explicit quality criteria described and applied", and one using "applied explicit quality criteria". However, these expressions are interpreted to convey the same meaning: the explicit definition and extraction of quality criteria from each reviewed study. This is because applying quality criteria inherently implies that the necessary data have been extracted from each reviewed study.

In the context of these 27 studies, 23 of them [10, 14, 16, 26, 28, 32- 37, 39, 40, 43-46, 50, 51, 54, 56, 57, 59] allocate a partial score when the research question involves addressing quality or validity issues. More precisely, as articulated in [56], "research questions from the secondary study address the quality of primary studies". Authors in [52] mention that a partial score is assigned if "the assessment is focused on answering the research questions posed in the primary paper". However, it is unclear how the quality assessment conducted as part of a secondary study, such as an SLR, contributes to answering questions in the primary paper. It is believed that there might be a mistake and that authors could be referring to the secondary paper instead. In addition, authors in [53] state that they assign a partial score if the quality assessment is conducted but not reported. However, it is unclear whether the quality assessment criteria or the results are reported or not. Authors in [47, 49], assign a partial score if the quality assessment is implicit. This entails various interpretations, including scenarios where predetermined criteria are not explicitly specified or a formalized process for assessment is lacking.

Authors in [38, 41, 42, 48] assigned a full score when the quality criteria were explicitly defined, without specifying anything about the application. In [38, 41, 42], a partial score was assigned when the quality criteria were assessed but not defined, whereas in [48], a partial score was assigned when the research questions involved quality issues addressed by the study. In [30], a full score was given when assessing quality without specifying anything about reporting quality criteria, while a partial score was given when only the study design was extracted but not assessed. In [55], the criteria for assigning full scores appear ambiguous, as the study simply stated that "the use of quality criteria is reported". A partial score was assigned if the research question mentioned quality aspects.

A score of 0 has been assigned in certain studies under the following circumstances related to quality assessment. In [38, 41, 42, 49, 53, 56], a score of 0 is assigned when there is no quality assessment. Furthermore, in [14, 28, 30, 32, 36, 40, 45, 46, 50-52, 54, 59], a score of 0 is assigned when there is no explicit quality assessment. The former condition implies a complete absence of quality assessment, while the latter suggests the presence of an assessment lacking transparency or specificity in terms of criteria, methods, or procedures. However, in the latter case, the distinction between assigning a partial score and no score becomes less clear, particularly concerning whether quality issues are addressed in the research questions and the absence of explicit quality assessment. Clarifying this difference is essential for a precise evaluation. Some studies, apart from lacking explicit quality assessment, include conditions in which quality data are extracted but not used [10, 33, 34, 37, 43, 44], defined but not used [35, 16, 26], or insufficiently described [57]. A score of 0 is assigned in [48], when there are no explicit quality criteria, and in [47], when no quality assessment criteria are defined or used. Additionally, authors in [39] follow a similar trend, assigning a score of 0 when there is an absence of a study quality assessment strategy, which it is believed that covers the absence of quality assessment criteria.

TABLE III. STUDIES EMPLOYING CUSTOM-MADE QUALITY INSTRUMENTS AND THE USED CRITERIA (RELATED TO RQ3)

Ref	Quality criteria
[17]	Is the publisher reputable?. What role has AI played in the review? E.g. primary technology under consideration, one of the two (or many) technologies considered. What type of review has been performed? Have the number and quality of primary studies been reported? How many online databases were searched? Are the years covered in the review known? Have specific SLR guidelines been reported to be followed in the review? Have the search strings been reported and how detailed they are in describing the AI? Has the data analysis method been described? Have the research questions been clearly defined?
[19]	RQ: Denotes whether the study mentions and answers the RQ or not. RQ Quality: Represents the quality of the RQ research in the secondary study. Future Prospects: A study that mentions the scope of future prospects or gives directions to contribute further to the area. Statistical Test: If statistical tests used in the area by different researchers were present in the secondary study. Tools available: If the tools used in the area of TCS&P were mentioned in the study. Detailed Analysis: A study that provided a detailed and in-depth analysis of the research work accomplished in the area of TCS&P. Novel Contribution: The study made a novel contribution in the analysis of the research conducted in TCS&P.
[22]	Does the selected study focus on the communication and coordination challenges or issues faced during the Requirements Change Management (RCM) stage? Does the selected study present any framework or model to overcome the communication and coordination challenges during RCM in Global Software Development (GSD) Does the selected study focus on GSD or distributed development context?
[23]	Are the experiences or findings drawn after actually conducting an SLR? Is the SLR from which the findings are drawn, sufficiently described to create the context and understanding for experience or findings? Are the experiences of other researchers consulted while describing their own? Are the experiences or findings linked clearly to one of the phases of SLR?
[24]	Do the research methods address the research questions? Does the study focus on code smells? Does the study discuss approaches, tools, and targeted domains? Are the data related to the topic? Are the results relevant to the research questions?
[25]	Does the SLR state how QA was performed? Does the SLR follow any QA instrument? Does the SLR state the purpose of performing QA? Does the SLR state what aspects were considered in QA? Does the SLR present the results of QA?

5) *Interpretation of the Quality Criteria related to the Information provided about Each Primary Study*

When assessing the information provided about the primary studies, attaining full scores in [14, 16, 26, 28, 32, 34-36, 39, 40, 45-47, 49, 50, 52, 53, 55, 56, 59] requires having information about each of the reviewed studies. While most studies use the terms "information" or "data" alone, some add adjectives and phrasing, such as "clear description of information", "complete information", "detailed information", or "relevant data" to convey the same meaning. Authors in [48] precisely outline the specific information required. Partial scores are assigned in 19 studies when only summary information about the primary studies is provided. It is important to note that in one study, the phrase "abstract level information" is used, which it is presumed that holds the same meaning as "summary information". Authors in [55] diverge in assigning a partial score based on the criterion of "human factors being grouped for analysis", specifically concerning the study's topic. Furthermore, in [10, 33, 37, 38, 41-44, 51, 57, 60], obtaining full scores requires detailed information about each paper, allowing individual papers to be traced back from data summaries. In other words, when papers are categorized, it should be possible to identify which individual studies belong to each category. This inference is drawn from the condition used in 7 studies for assigning partial scores, indicating that only summary information is presented, and while papers are grouped into categories, linking individual studies to these categories is not possible. The remaining 3 studies assign partial scores when presenting summary information.

A score of 0 is assigned when the results or information about primary studies are not specified. Some authors further elaborate that this means that individual studies are not being cited. Other studies use different expressions, such as "information presented is not referenced", "primary studies are not detailed", and "outcomes of specific studies are not quantified". However, it remains unclear how this is distinguished from presenting only summary information, which is the basis for assigning a partial score. Notably, in [48, 51], a 0 score is explicitly assigned when "results on individual studies are neither specified nor summarized", and there is "no information (specific or generalized) about the included sources". In [30], a full score is achieved by providing a reference to each study, which appears synonymous with the ability to trace back individual studies. A partial score is granted when only the number of studies is provided, whereas a score of 0 is assigned if even the number is not provided.

IV. DISCUSSION

A. Basic Information

The emphasis on software engineering is driven by historical factors, as SLRs in computing were initially introduced within software engineering. SLR rapid adoption and popularity occurred swiftly in this field, resulting in a substantial volume of systematic review studies. Consequently, this facilitates the conduct of tertiary studies within this specific field of computing.

B. Current Instruments

A widespread acceptance of the DARE framework for evaluating quality was observed. Two factors underlie the popularity of DARE over other quality instruments. Firstly, it consists of fewer items compared to alternative instruments such as PRISMA, which has 27 items, and AMSTAR, which has 11 items [61]. Therefore, it is assumed that DARE is perceived as easier and less time-consuming to apply. The second factor contributing to DARE's popularity is its adoption by Kitchenham et al. in two seminal tertiary studies [10, 14]. The widespread adoption of DARE in subsequent studies may be attributed to the influence of Kitchenham, a key figure renowned for introducing SLRs in software engineering and providing guidelines for their conduction [62]. As of June 12, 2024, this work has received 12,439 citations according to Google Scholar.

Despite its widespread use, it is essential to acknowledge the recognized limitations of the DARE framework. Its simplicity results in gaps in covering various quality issues and various phases of SLRs. For instance, it does not include an assessment of threats to validity, which is expected to be discussed in secondary studies [12]. Another limitation is that it focuses merely on determining whether specific activities within SLRs have been performed, rather than assessing the quality or effectiveness of their execution [50]. Even when studies incorporate the refinements described in [14], with certain criteria rephrased to emphasize rigor and go beyond mere reporting, these refinements are not adequately accounted for in the interpretation. For instance, whether the question is phrased as "were inclusion/exclusion criteria reported?" or "are the review's inclusion and exclusion criteria described and appropriate?", the interpretations remain largely consistent.

C. Modification of Quality Instruments

The aforementioned limitations have prompted some studies to customize the discussed framework, addressing missing aspects or adapting it to their specific contexts [12]. However, the modifications to DARE were minimal, with only six studies having introduced changes by adding new criteria. This further supports this study's claim that its popularity is attributable to its simplicity. The present work supports the proposal made in [12] regarding the development of a comprehensive framework that encompasses all SLR phases while providing authors with the flexibility to customize it by excluding criteria that do not apply to their study. This approach minimizes the need for authors to independently extend the framework, thereby fostering consistency. As proposed in [12], one could start with DARE and incorporate any missing elements. Additionally, other available frameworks should be evaluated and combined to cover all phases and issues of SLRs.

D. Interpretation of Quality Instruments

The performed analysis of interpretations reveals variations in the conditions used by the studies to assign scores for each DARE criterion. A common limitation among them is the lack of traceability linking the proposed interpretations with the evidence or best practices that motivated them [2]. It is proposed that authors provide examples and justify any changes made, as this practice ensures a clearer understanding of the limitations associated with the existing interpretations. This includes offering explicit instructions on applying the criteria, interpreting the results, and addressing any ambiguities that may arise. Authors should also highlight the difficulties faced during the assessment, as this can contribute to the refinement of the framework or interpretations. A lack of measurable metrics in the interpretation is also observed. It is crucial to operationalize phrases, such as "all journals addressing the topic of interest", "restricted set of journals", "extremely restricted set of journals", and "reputable database". Operationalizing the scoring process is crucial to ensure consistency, reliability, and comparability across diverse studies. The first step in this direction involves, for example, providing definitions for what constitutes a "reputable database" in each study. These definitions should be then collaboratively combined based on studies within the same topics or areas.

The following sections explore whether and how the common interpretations identified in this analysis could be automated, to reduce subjectivity and enhance the reproducibility of assessments.

1) Interpretation of the Quality Criteria related to Inclusion/Exclusion Criteria

Automation is feasible for certain aspects of the assessment of inclusion and exclusion criteria in SLRs. Text Mining (TM) techniques could identify whether inclusion or exclusion criteria are explicitly stated by scanning for keywords, section headings, or specific structural markers that outline these criteria. Detecting implicit criteria is far more challenging, as it requires a contextual understanding. Implicit criteria may include participant characteristics (e.g., age, gender) or methodological constraints (e.g., sample size, time frame) that are not explicitly stated as inclusion or exclusion criteria but can be inferred from the research question or title. Named Entity Recognition (NER) models can identify these key entities, enabling the detection of implicit criteria. Assessing the appropriateness of criteria is also challenging, as it is highly context-dependent and often subjective, relying on domain knowledge and an understanding of the research objectives. However, automated assessments could focus on identifying justifications, especially for inclusion and exclusion criteria, to flag instances where criteria may lack sufficient rationale. This would provide a basis for partial automation, highlighting areas that need further human review.

2) Interpretation of the Quality Criteria related to the Adequacy of the Search

Based on the analysis of interpretations related to the quality criteria for search adequacy, automation seems to be a viable option for evaluating the adequacy of the search strategies in SLRs. TM techniques can be leveraged to detect descriptions of search strategies within the SLR method sections. Since score assignment often depends on the number of search strategies and the use of specific strategies, automation could potentially manage these aspects.

Authors in [63] have already explored automation techniques for identifying and extracting information about the referenced databases in the considered studies. Building on this work, automated systems could be further developed to not only detect which databases are used in SLRs, but also to assess their appropriateness in relation to the specific research field. While the assessment of search adequacy typically focuses on the number of databases referenced in a review, automation could go beyond this by evaluating whether the selected databases are appropriate for the research topic. For instance, an automated system could cross-reference the databases used in a given review with a curated list of commonly accepted or highly relevant databases in the field. In health-related research, for example, databases like PubMed or the Cochrane Library are often considered central to the field, whereas in engineering, databases such as IEEE Xplore or Compendex are highly relevant. This could help identify potential gaps in the review's search strategy, such as overlooking important field-specific databases. Additionally, it is possible to operationalize and assess the concept of "all relevant journals" within a given subject area. By using citation analysis and bibliometric methods, it is possible to compile a list of journals that are considered highly relevant to the study area. These journals could then serve as a benchmark for assessing the fulfillment of the criterion of "identifying and searching all relevant journals on the subject of interest". This is performed by checking whether these journals are adequately represented in reviews within the same subject area.

3) Interpretation of the Quality Criteria related to the Synthesis Method

Evaluating the quality criteria associated with the synthesis methods used in SLRs can be automated. In the most common scenario, where a full score is awarded when the synthesis method is named and referenced, TM can extract the mentions of the synthesis method from the text. Existing research, such as [64], has explored the use of TM techniques to extract methodological information. These methods can be adapted to specifically identify the synthesis method employed in SLRs. TM can then search for citation formats (e.g. "(Author, Year)" or "Author [Year]") and verify whether these references correspond to the specified synthesis method.

4) Interpretation of the Quality Criteria related to the Quality/Validity

The identification of quality criteria in SLRs is increasingly achievable through a combination of TM and Machine Learning (ML) techniques. TM can extract quality criteria by using specific keywords, phrases, and contextual clues. Key quality criteria typically include sample size, methodological rigor, statistical significance, bias, ethical considerations, and replication or reproducibility. By applying ML models, sentences or paragraphs within SLRs can be automatically classified according to their relevance to these criteria. These models can be trained to differentiate between various aspects of research quality, such as identifying whether passages pertain to methodological descriptions or report on the outcomes of quality assessments. The process of assessing whether a research question addresses quality or validity issues can also be automated. This begins with extracting research questions from the text and classifying them based on the identification of relevant keywords or phrases that signal concerns related to quality or validity, using classification techniques [113]. By analyzing these indicators, the system can determine whether a question is related to issues, such as methodological rigor, reliability, potential biases, or other quality aspects.

5) Interpretation of the Quality Criteria related to the Information provided about Each Primary Study

Assessing whether the necessary information is available for each reviewed study can be challenging, but it is achievable through the application of TM techniques. The first step in automating this process is to identify the specific data elements that have been extracted, which are typically outlined in the review methodology section. Once the target data elements are defined, the system should then check whether all required data elements (as defined in the review methodology) are present for each study. Common data elements might include study design and methodology, sample size, participant demographics, and primary outcomes or results. Although some information might be directly described in the text body, other data may appear in supplementary formats, like tables, figures, or appendices. An automated system should account for extracting data from these different sources. Additionally, the system should validate the completeness of the data by checking whether all required elements are available for every study. If any data points are missing or incomplete, the system can flag these studies for further review.

Automatically recognizing whether only summary information (e.g., aggregated or generalized data) is present can be achieved using ML models. These models can be trained to distinguish between summary data, which typically include high-level overviews, trends, or aggregated results, and detailed data, which contain specific, precise values, such as numerical figures or other empirical findings. By analyzing patterns in the text language and structure, ML models can identify the presence of detailed data and flag sections that contain only summaries, helping to automate the assessment of these quality criteria.

V. CONCLUSION

One of the primary motivations driving this study was to gain a deeper understanding of the quality assessment practices specifically applied to the evaluation of Systematic Literature Reviews (SLRs) in computing, laying the groundwork for automation. The study addresses key questions about the current instruments, their usage, modification, and how researchers interpret the quality criteria. It contributes by highlighting issues related to diverse interpretations and potential inconsistencies in the application of quality instruments and by providing actionable recommendations for automating the quality assessment process, thus contributing to the ongoing efforts to enhance SLR quality assessment.

To facilitate the transition to automation, several steps have been proposed including the development of a comprehensive quality framework, building on DARE and other existing frameworks, the refinement and operationalization of the scoring process (by defining measurable versions), and the application of Text Mining (TM) and Machine Learning (ML) techniques to automate the scoring of quality criteria.

Future research should focus on developing automated tools that support researchers in applying quality assessment frameworks to their reviews or evaluating other reviews. This could enable the research community to adopt standardized practices for SLRs more easily. It is essential to combine automation with human oversight in these tools, as although automation can handle tasks such as data extraction and preliminary analysis, human reviewers should provide the final judgments.

Finally, it is believed that automating quality assessment will serve as a cornerstone in elevating the overall process quality. Automation will reduce manual effort and enable the use of more comprehensive tools, creating universal rules that minimize inconsistencies in study interpretations. However, while automation can support the quality assessment process, it cannot completely replace human judgment. The most effective approach will likely involve a hybrid system, where automation handles data extraction and analysis, while human reviewers focus on higher-level decision-making.

DATA AVAILABILITY

The complete extracted data are available at:

https://docs.google.com/document/d/1gr0e49aLJctSDEWDvaymRTWpS-8fIay6BK6Cy_4-vHQ/edit?usp=sharing.

REFERENCES

[1] B. A. Kitchenham, "Systematic reviews," in *10th International Symposium on Software Metrics, 2004. Proceedings.*, Sep. 2004, pp. xii–xii, https://doi.org/10.1109/METRIC.2004.1357885.

[2] M. Usman, N. bin Ali, and C. Wohlin, "A Quality Assessment Instrument for Systematic Literature Reviews in Software Engineering," *e-Informatica Software Engineering Journal*, vol. 17, no. 1, 2023, Art. no. 230105, https://doi.org/10.37190/e-Inf230105.

[3] B. Kitchenham, L. Madeyski, and D. Budgen, "SEGRESS: Software Engineering Guidelines for REporting Secondary Studies," *IEEE Transactions on Software Engineering*, vol. 49, no. 3, pp. 1273–1298, Mar. 2023, https://doi.org/10.1109/TSE.2022.3174092.

[4] Z. Hamad and N. Salim, "Systematic literature review (SLR) automation: A systematic literature review," *Journal of Theoretical and Applied Information Technology*, vol. 59, no. 3, pp. 661–672, 2014.

[5] N. bin Ali and M. Usman, "A critical appraisal tool for systematic literature reviews in software engineering," *Information and Software Technology*, vol. 112, pp. 48–50, Aug. 2019, https://doi.org/10.1016/j.infsof.2019.04.006.

[6] D. Budgen, P. Brereton, S. Drummond, and N. Williams, "Reporting systematic reviews: Some lessons from a tertiary study," *Information and Software Technology*, vol. 95, pp. 62–74, Mar. 2018, https://doi.org/10.1016/j.infsof.2017.10.017.

[7] 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, pp. 133–147, Jul. 2018, https://doi.org/10.1016/j.infsof.2018.02.002.

[8] 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*, Hamilton, New Zealand, Dec. 2016, pp. 153–160, https://doi.org/10.1109/APSEC.2016.031.

[9] J. Kruger, C. Lausberger, I. von Nostitz-Wallwitz, G. Saake, and T. Leich, "Search. Review. Repeat? An empirical study of threats to replicating SLR searches," *Empirical Software Engineering*, vol. 25, no. 1, pp. 627–677, Jan. 2020, https://doi.org/10.1007/s10664-019-09763-0.

[10] B. Kitchenham *et al.*, "Systematic literature reviews in software engineering – A tertiary study," *Information and Software Technology*, vol. 52, no. 8, pp. 792–805, Aug. 2010, https://doi.org/10.1016/j.infsof.2010.03.006.

[11] F. Q. B. da Silva, A. L. M. Santos, S. Soares, A. C. C. Franca, and C. V. F. Monteiro, "Six Years of Systematic Literature Reviews in Software Engineering: an Extended Tertiary Study," in *International Conference on Software Engineering*, Cape Town, South Africa, Dec. 2010, pp. 1–10.

[12] D. Costal, C. Farre, X. Franch, and C. Quer, "How Tertiary Studies perform Quality Assessment of Secondary Studies in Software Engineering." arXiv, Oct. 07, 2021, https://doi.org/10.48550/arXiv.2110.03820.

[13] G. Kurdi, "Building an Electronic Resource of Systematic Reviews in Computing," in *IEEE International Conference on Computing*, Langkawi, Malaysia, Oct. 2023, pp. 95–100, https://doi.org/10.1109/ICOCO59262.2023.10397762.

[14] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, https://doi.org/10.1016/j.infsof.2008.09.009.

[15] F. K. Kamei, "The Use of Grey Literature Review as Evidence for Practitioners," *ACM SIGSOFT Software Engineering Notes*, vol. 44, no. 3, Jul. 2020, Art. no. 23, https://doi.org/10.1145/3356773.3356797.

[16] J. M. Verner, O. P. Brereton, B. A. Kitchenham, M. Turner, and M. Niazi, "Risks and risk mitigation in global software development: A tertiary study," *Information and Software Technology*, vol. 56, no. 1, pp. 54–78, Jan. 2014, https://doi.org/10.1016/j.infsof.2013.06.005.

[17] M. Cubric, "Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study," *Technology in Society*, vol. 62, Aug. 2020, Art. no. 101257, https://doi.org/10.1016/j.techsoc.2020.101257.

[18] R. Kachouie, S. Sedighadeli, and A. B. Abkenar, "The Role of Socially Assistive Robots in Elderly Wellbeing: A Systematic Review," in *International Conference on Human-Computer Interaction*, Vancouver, BC, Canada, Jul. 2017, pp. 669–682, https://doi.org/10.1007/978-3-319-57931-3_54.

[19] S. Singhal, N. Jatana, B. Suri, S. Misra, and L. Fernandez-Sanz, "Systematic Literature Review on Test Case Selection and Prioritization: A Tertiary Study," *Applied Sciences*, vol. 11, no. 24, Jan. 2021, Art. no. 12121, https://doi.org/10.3390/app112412121.

[20] E. Engstrom, P. Runeson, and M. Skoglund, "A systematic review on regression test selection techniques," *Information and Software Technology*, vol. 52, no. 1, pp. 14–30, Jan. 2010, https://doi.org/10.1016/j.infsof.2009.07.001.

[21] B. Suri and S. Singhal, "Implementing Ant Colony Optimization for Test Case Selection and Prioritization," *International Journal on Computer Science and Engineering*, vol. 3, no. 5, pp. 1924–1932, 2011.

[22] S. Qureshi, S. U. R. Khan, J. Iqbal, and Inayat-Ur-Rehman, "A Study on Mitigating the Communication and Coordination Challenges During Requirements Change Management in Global Software Development,"

IEEE Access, vol. 9, pp. 88217–88242, Jan. 2021, https://doi.org/10.1109/ACCESS.2021.3090098.

[23] S. Imtiaz, M. Bano, N. Ikram, and M. Niazi, "A tertiary study: experiences of conducting systematic literature reviews in software engineering," in *17th International Conference on Evaluation and Assessment in Software Engineering*, Porto de Galinhas, Brazil, Apr. 2013, pp. 177–182, https://doi.org/10.1145/2460999.2461025.

[24] R. Yaqoob, Sanaa, S. U. R. Khan, and M. A. Shah, "Tertiary study on landscaping the review in code smells," *IET Conference Proceedings*, vol. 2021, no. 4, pp. 131–136, Oct. 2021, https://doi.org/10.1049/icp.2021.2421.

[25] L. Yang *et al.*, "Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective," *Information and Software Technology*, vol. 130, Feb. 2021, Art. no. 106397, https://doi.org/10.1016/j.infsof.2020.106397.

[26] P. Malcher, O. Barbosa, D. Viana, and R. Santos, "Software Ecosystems: A Tertiary Study and a Thematic Model." arXiv, Dec. 20, 2022, https://doi.org/10.48550/arXiv.2212.10443.

[27] P. A. S. Duarte, F. M. Barreto, P. A. C. Aguilar, J. Boudy, R. M. C. Andrade, and W. Viana, "AAL Platforms Challenges in IoT Era: A Tertiary Study," in *13th Annual Conference on System of Systems Engineering*, Paris, France, Jun. 2018, pp. 106–113, https://doi.org/10.1109/SYSOSE.2018.8428745.

[28] J. L. Barros-Justo, F. B. V. Benitti, and S. Matalonga, "Trends in software reuse research: A tertiary study," *Computer Standards & Interfaces*, vol. 66, Oct. 2019, Art. no. 103352, https://doi.org/10.1016/j.csi.2019.04.011.

[29] Q. Xu, X. Chen, S. Li, H. Zhang, M. A. Babar, and N. K. Tran, "Blockchain-based Solutions for IoT: A Tertiary Study," in *20th International Conference on Software Quality, Reliability and Security Companion*, Macau, China, Dec. 2020, pp. 124–131, https://doi.org/10.1109/QRS-C51114.2020.00031.

[30] M. Raatikainen, J. Tiihonen, and T. Mannisto, "Software product lines and variability modeling: A tertiary study," *Journal of Systems and Software*, vol. 149, pp. 485–510, Mar. 2019, https://doi.org/10.1016/j.jss.2018.12.027.

[31] S. P. Pillai, S. D. Madhukumar, and T. Radharamanan, "Consolidating evidence based studies in software cost/effort estimation — A tertiary study," in *TENCON 2017 - 2017 IEEE Region 10 Conference*, Penang, Malaysia, Nov. 2017, pp. 833–838, https://doi.org/10.1109/TENCON.2017.8227974.

[32] A. Idri and L. Cheikhi, "A survey of secondary studies in software process improvement," in *13th International Conference of Computer Systems and Applications*, Agadir, Morocco, Dec. 2016, pp. 1–8, https://doi.org/10.1109/AICCSA.2016.7945655.

[33] F. Q. B. da Silva, A. L. M. Santos, S. Soares, A. C. C. França, C. V. F. Monteiro, and F. F. Maciel, "Six years of systematic literature reviews in software engineering: An updated tertiary study," *Information and Software Technology*, vol. 53, no. 9, pp. 899–913, Sep. 2011, https://doi.org/10.1016/j.infsof.2011.04.004.

[34] A. B. Marques, R. Rodrigues, and T. Conte, "Systematic Literature Reviews in Distributed Software Development: A Tertiary Study," in *Seventh International Conference on Global Software Engineering*, Porto Alegre, Brazil, Aug. 2012, pp. 134–143, https://doi.org/10.1109/ICGSE.2012.29.

[35] J. M. Verner, O. P. Brereton, B. A . Kitchenham, M. Turner, and M. Niazi, "Systematic literature reviews in global software development: A tertiary study," in *16th International Conference on Evaluation & Assessment in Software Engineering*, Ciudad Real, Spain, Dec. 2012, pp. 2–11, https://doi.org/10.1049/ic.2012.0001.

[36] C. Marimuthu and K. Chandrasekaran, "Systematic Studies in Software Product Lines: A Tertiary Study," in *21st International Systems and Software Product Line Conference*, Sevilla, Spain, Sep. 2017, pp. 143–152, https://doi.org/10.1145/3106195.3106212.

[37] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 4997–5053, Oct. 2021, https://doi.org/10.1007/s10462-021-09973-3.

[38] S. Saleem, N. F. Khan, S. Zafar, and N. Raza, "Systematic literature reviews in cyberbullying/cyber harassment: A tertiary study," *Technology in Society*, vol. 70, Aug. 2022, Art. no. 102055, https://doi.org/10.1016/j.techsoc.2022.102055.

[39] A. A. Khan, J. Keung, M. Niazi, S. Hussain, and H. Zhang, "Systematic Literature Reviews of Software Process Improvement: A Tertiary Study," in *European Conference on Software Process Improvement*, Ostrava, Czech Republic, Sep. 2017, pp. 177–190, https://doi.org/10.1007/978-3-319-64218-5_14.

[40] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli, "Usability in agile software development: A tertiary study," *Computer Standards & Interfaces*, vol. 64, pp. 61–77, May 2019, https://doi.org/10.1016/j.csi.2018.12.003.

[41] N. F. Khan, A. Yaqoob, M. S. Khan, and N. Ikram, "The cybersecurity behavioral research: A tertiary study," *Computers & Security*, vol. 120, Sep. 2022, Art. no. 102826, https://doi.org/10.1016/j.cose.2022.102826.

[42] R. Hoda, N. Salleh, J. Grundy, and H. M. Tee, "Systematic literature reviews in agile software development: A tertiary study," *Information and Software Technology*, vol. 85, pp. 60–70, May 2017, https://doi.org/10.1016/j.infsof.2017.01.007.

[43] L. Villalobos Arias, C. U. Quesada López, A. Martínez Porras, and M. Jenkins Coronas, "A tertiary study on model-based testing areas, tools and challenges: Preliminary results," 2018.

[44] L. Villalobos-Arias, C. Quesada-Lopez, A. Martinez, and M. Jenkins, "Model-based testing areas, tools and challenges: A tertiary study," *CLEI Electronic Journal*, vol. 22, no. 1, Apr. 2019, Art. no. 3, https://doi.org/10.19153/cleiej.22.1.3.

[45] N. Rios, M. G. de Mendonça Neto, and R. O. Spinola, "A tertiary study on technical debt: Types, management strategies, research trends, and base information for practitioners," *Information and Software Technology*, vol. 102, pp. 117–145, Oct. 2018, https://doi.org/10.1016/j.infsof.2018.05.010.

[46] J. W. M. Lai and M. Bower, "Evaluation of technology use in education: Findings from a critical analysis of systematic literature reviews," *Journal of Computer Assisted Learning*, vol. 36, no. 3, pp. 241–259, 2020, https://doi.org/10.1111/jcal.12412.

[47] M. Bano, D. Zowghi, and N. Ikram, "Systematic reviews in requirements engineering: A tertiary study," in *4th International Workshop on Empirical Requirements Engineering*, Karlskrona, Sweden, Aug. 2014, pp. 9–16, https://doi.org/10.1109/EmpiRE.2014.6890110.

[48] H. J. Junior and G. H. Travassos, "Consolidating a common perspective on *Technical Debt* and its Management through a Tertiary Study," *Information and Software Technology*, vol. 149, Sep. 2022, Art. no. 106964, https://doi.org/10.1016/j.infsof.2022.106964.

[49] Z. Kotti, R. Galanopoulou, and D. Spinellis, "Machine Learning for Software Engineering: A Tertiary Study," *ACM Comput. Surv.*, vol. 55, no. 12, Nov. 2023, Art. no. 256, https://doi.org/10.1145/3572905.

[50] D. Budgen, P. Brereton, N. Williams, and S. Drummond, "What Support do Systematic Reviews Provide for Evidence-informed Teaching about Software Engineering Practice?," *e-Informatica Software Engineering Journal*, vol. 14, no. 1, pp. 7–60, 2020, https://doi.org/10.37190/e-Inf200101.

[51] H. Cadavid, V. Andrikopoulos, and P. Avgeriou, "Architecting systems of systems: A tertiary study," *Information and Software Technology*, vol. 118, Feb. 2020, Art. no. 106202, https://doi.org/10.1016/j.infsof.2019.106202.

[52] G. A. Garcia-Mireles and M. E. Morales-Trujillo, "Gamification in Software Engineering: A Tertiary Study," in *International Conference on Software Process Improvement*, Guanajuato, Mexico, Oct. 2019, pp. 116–128, https://doi.org/10.1007/978-3-030-33547-2_10.

[53] 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, pp. 396–436, Mar. 2019, https://doi.org/10.1016/j.jss.2018.12.018.

[54] M. Goulao, V. Amaral, and M. Mernik, "Quality in model-driven engineering: a tertiary study," *Software Quality Journal*, vol. 24, no. 3, pp. 601–633, Sep. 2016, https://doi.org/10.1007/s11219-016-9324-8.

[55] E. Dutra, B. Diirr, and G. Santos, "Human Factors and their Influence on Software Development Teams - A Tertiary Study," in *XXXV Brazilian Symposium on Software Engineering*, Joinville, Brazil, Oct. 2021, pp. 442–451, https://doi.org/10.1145/3474624.3474625.

[56] T. N. Kudo, R. F. Bulcao-Neto, and A. M. R. Vincenzi, "Requirement patterns: a tertiary study and a research agenda," *IET Software*, vol. 14, no. 1, pp. 18–26, 2020, https://doi.org/10.1049/iet-sen.2019.0016.

[57] H. K. V. Tran, M. Unterkalmsteiner, J. Borstler, and N. bin Ali, "Assessing test artifact quality—A tertiary study," *Information and Software Technology*, vol. 139, Nov. 2021, Art. no. 106620, https://doi.org/10.1016/j.infsof.2021.106620.

[58] S. Das and K. Gary, "Agile Transformation at Scale: A Tertiary Study," in *International Conference on Agile Software Development*, Jun. 2021, pp. 3–11, https://doi.org/10.1007/978-3-030-88583-0_1.

[59] D. Budgen, S. Drummond, P. Brereton, and N. Holland, "What scope is there for adopting evidence-informed teaching in SE?," in *34th International Conference on Software Engineering*, Zurich, Switzerland, Jun. 2012, pp. 1205–1214, https://doi.org/10.1109/ICSE.2012.6227022.

[60] V. Garousi and M. V. Mantyla, "A systematic literature review of literature reviews in software testing," *Information and Software Technology*, vol. 80, pp. 195–216, Dec. 2016, https://doi.org/10.1016/j.infsof.2016.09.002.

[61] B. J. Shea *et al.*, "AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews," *Journal of Clinical Epidemiology*, vol. 62, no. 10, pp. 1013–1020, Oct. 2009, https://doi.org/10.1016/j.jclinepi.2008.10.009.

[62] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," University of Durham, Durham, UK, Technical Report EBSE-2007-01, Jan. 2007.

[63] G. Duck, G. Nenadic, A. Brass, D. L. Robertson, and R. Stevens, "Extracting patterns of database and software usage from the bioinformatics literature," *Bioinformatics*, vol. 30, no. 17, pp. i601–i608, Sep. 2014, https://doi.org/10.1093/bioinformatics/btu471.

[64] A. Kovacevic, Z. Konjovic, B. Milosavljevic, and G. Nenadic, "Mining methodologies from NLP publications: A case study in automatic terminology recognition," *Computer Speech & Language*, vol. 26, no. 2, pp. 105–126, Apr. 2012, https://doi.org/10.1016/j.csl.2011.09.001.

[65] D. S. Cruzes and T. Dyba, "Synthesizing evidence in software engineering research," in *ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, Bolzano-Bozen, Italy, Sep. 2010, pp. 1–10, https://doi.org/10.1145/1852786.1852788.

[66] F. Lau, C. Kuziemsky, M. Price, and J. Gardner, "A review on systematic reviews of health information system studies," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 637–645, Nov. 2010, https://doi.org/10.1136/jamia.2010.004838.

[67] D. S. Cruzes and T. Dyba, "Research synthesis in software engineering: A tertiary study," *Information and Software Technology*, vol. 53, no. 5, pp. 440–455, May 2011, https://doi.org/10.1016/j.infsof.2011.01.004.

[68] A. C. C. dos Santos, I. H. de F. Junior, H. P. de Moura, and S. Marczak, "A Systematic Tertiary Study of Communication in Distributed Software Development Projects," in *Seventh International Conference on Global Software Engineering*, Porto Alegre, Brazil, Aug. 2012, pp. 182–182, https://doi.org/10.1109/ICGSE.2012.42.

[69] 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, pp. 162–183, Sep. 2016, https://doi.org/10.1016/j.jss.2016.06.043.

[70] Z. I. Batouta, R. Dehbi, M. Talea, and O. Hajoui, "Automation in code generation: Tertiary and systematic mapping review," in *4th IEEE International Colloquium on Information Science and Technology*, Tangier, Morocco, Oct. 2016, pp. 200–205, https://doi.org/10.1109/CIST.2016.7805042.

[71] E. Oliveira, T. Conte, M. Cristo, and N. Valentim, "Influence Factors in Software Productivity — A Tertiary Literature Review," *International Journal of Software Engineering and Knowledge Engineering*, vol. 28, no. 11n12, pp. 1795–1810, Nov. 2018, https://doi.org/10.1142/S0218194018400296.

[72] 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, pp. 201–230, Feb. 2019, https://doi.org/10.1016/j.infsof.2018.10.006.

[73] L. Bedu, O. Tinh, and F. Petrillo, "A Tertiary Systematic Literature Review on Software Visualization," in *Working Conference on Software Visualization*, Cleveland, OH, USA, Oct. 2019, pp. 33–44, https://doi.org/10.1109/VISSOFT.2019.00013.

[74] V. Delavari, E. Shaban, M. Janssen, and A. Hassanzadeh, "Thematic mapping of cloud computing based on a systematic review: a tertiary study," *Journal of Enterprise Information Management*, vol. 33, no. 1, pp. 161–190, Sep. 2019, https://doi.org/10.1108/JEIM-02-2019-0034.

[75] G. T. G. Neto, W. B. Santos, P. T. Endo, and R. A. A. Fagundes, "Multivocal literature reviews in software engineering: Preliminary findings from a tertiary study," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, Porto de Galinhas, Brazil, Sep. 2019, pp. 1–6, https://doi.org/10.1109/ESEM.2019.8870142.

[76] D. Akdur and O. Demirors, "Systematic Reviews in Model-Driven Engineering: a Tertiary Study," *Journal of Aeronautics and Space Technologies*, vol. 13, no. 1, pp. 57–68, 2020.

[77] J. L. Barros-Justo, F. B. V. Benitti, and J. S. Molleri, "Risks and risk mitigation in global software development: An update," *Journal of Software: Evolution and Process*, vol. 33, no. 11, 2021, Art. no. e2370, https://doi.org/10.1002/smr.2370.

[78] X. Huang, H. Zhang, X. Zhou, D. Shao, and L. Jaccheri, "A Research Landscape of Software Engineering Education," in *28th Asia-Pacific Software Engineering Conference*, Taipei, Taiwan, Dec. 2021, pp. 181–191, https://doi.org/10.1109/APSEC53868.2021.00026.

[79] E. M. Arvanitou, A. Ampatzoglou, S. Bibi, A. Chatzigeorgiou, and I. Deligiannis, "Applying and Researching DevOps: A Tertiary Study," *IEEE Access*, vol. 10, pp. 61585–61600, Jan. 2022, https://doi.org/10.1109/ACCESS.2022.3171803.

[80] H. Y. Fuchigami, M. R. Severino, L. Yamanaka, and M. R. de Oliveira, "A Literature Review of Mathematical Programming Applications in the Fresh Agri-Food Supply Chain," in *International Joint conference on Industrial Engineering and Operations Management*, Novi Sad, Serbia, Jul. 2019, pp. 37–50, https://doi.org/10.1007/978-3-030-14973-4_4.

[81] F. Kamei, G. Pinto, I. Wiese, M. Ribeiro, and S. Soares, "What Evidence We Would Miss If We Do Not Use Grey Literature?," in *15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement*, Bari, Italy, Oct. 2021, pp. 1–11, https://doi.org/10.1145/3475716.3475777.

[82] N. Salleh and A. Nordin, "Trends and perceptions of evidence-based software engineering research in Malaysia," in *The 5th International Conference on Information and Communication Technology for The Muslim World*, Kuching, Malaysia, Nov. 2014, pp. 1–6, https://doi.org/10.1109/ICT4M.2014.7020597.

[83] M. Neumann, "The Integrated List of Agile Practices - A Tertiary Study," in *International Conference on Lean and Agile Software Development*, Jan. 2022, pp. 19–37, https://doi.org/10.1007/978-3-030-94238-0_2.

[84] J. Barata, "The fourth industrial revolution of supply chains: A tertiary study," *Journal of Engineering and Technology Management*, vol. 60, Apr. 2021, Art. no. 101624, https://doi.org/10.1016/j.jengtecman.2021.101624.

[85] D. Budgen, P. Brereton, N. Williams, and S. Drummond, "The contribution that empirical studies performed in industry make to the findings of systematic reviews: A tertiary study," *Information and Software Technology*, vol. 94, pp. 234–244, Feb. 2018, https://doi.org/10.1016/j.infsof.2017.10.012.

[86] C. Fu, H. Zhang, X. Huang, X. Zhou, and Z. Li, "A Review of Meta-ethnographies in Software Engineering," in *23rd International Conference on Evaluation and Assessment in Software Engineering*, Copenhagen, Denmark, Apr. 2019, pp. 68–77, https://doi.org/10.1145/3319008.3319015.

[87] D. Srivastava and V. Kumar Soni, "A Systematic Review On Sentiment Analysis Approaches," in *2nd International Conference on Advance*

Computing and Innovative Technologies in Engineering, Greater Noida, India, Apr. 2022, pp. 1–6, https://doi.org/10.1109/ICACITE53722.2022.9823769.

[88] H. Tang, Y. Zhou, X. Huang, and G. Rong, "Does pareto's law apply to evidence distribution in software engineering? an initial report," in *3rd International Workshop on Evidential Assessment of Software Technologies*, Nanjing, China, Dec. 2014, pp. 9–16, https://doi.org/10.1145/2627508.2627510.

[89] F. Kamei *et al.*, "Grey Literature in Software Engineering: A critical review," *Information and Software Technology*, vol. 138, Oct. 2021, Art. no. 106609, https://doi.org/10.1016/j.infsof.2021.106609.

[90] P. Singh, M. Galster, and K. Singh, "How do Secondary Studies in Software Engineering report Automated Searches? A Preliminary Analysis," in *22nd International Conference on Evaluation and Assessment in Software Engineering*, Christchurch, New Zealand, Jun. 2018, pp. 145–150, https://doi.org/10.1145/3210459.3210474.

[91] D. L. Cortes-Murcia, W. J. Guerrero, and J. R. Montoya-Torres, "Supply Chain Management, Game-Changing Technologies, and Physical Internet: A Systematic Meta-Review of Literature," *IEEE Access*, vol. 10, pp. 61721–61743, Jan. 2022, https://doi.org/10.1109/ACCESS.2022.3181154.

[92] G. K. Hanssen, D. Smite, and N. B. Moe, "Signs of Agile Trends in Global Software Engineering Research: A Tertiary Study," in *Sixth International Conference on Global Software Engineering Workshop*, Helsinki, Finland, Aug. 2011, pp. 17–23, https://doi.org/10.1109/ICGSE-W.2011.12.

[93] J. C. Epifanio and L. F. Da Silva, "Scrutinizing Reviews on Computer Science Technologies for Autism: Issues and Challenges," *IEEE Access*, vol. 8, pp. 32802–32815, Jan. 2020, https://doi.org/10.1109/ACCESS.2020.2973097.

[94] E. Bayram, B. Dogan, and V. Tunali, "Bibliometric Analysis of the Tertiary Study on Agile Software Development using Social Network Analysis," in *Innovations in Intelligent Systems and Applications Conference*, Istanbul, Turkey, Oct. 2020, pp. 1–4, https://doi.org/10.1109/ASYU50717.2020.9259875.

[95] S. Jiang, K. Jakobsen, L. Jaccheri, and J. Li, "Blockchain and Sustainability: A Tertiary Study," in *IEEE/ACM International Workshop on Body of Knowledge for Software Sustainability*, Madrid, Spain, Jun. 2021, pp. 7–8, https://doi.org/10.1109/BoKSS52540.2021.00011.

[96] H. Abedinnia, C. H. Glock, E. H. Grosse, and M. Schneider, "Machine scheduling problems in production: A tertiary study," *Computers & Industrial Engineering*, vol. 111, pp. 403–416, Sep. 2017, https://doi.org/10.1016/j.cie.2017.06.026.

[97] M. A. M. S. Lemstra and M. A. de Mesquita, "Industry 4.0: a tertiary literature review," *Technological Forecasting and Social Change*, vol. 186, Jan. 2023, Art. no. 122204, https://doi.org/10.1016/j.techfore.2022.122204.

[98] B. Cartaxo, "Integrating evidence from systematic reviews with software engineering practice through evidence briefings," in *20th International Conference on Evaluation and Assessment in Software Engineering*, Limerick, Ireland, Jun. 2016, pp. 1–4, https://doi.org/10.1145/2915970.2915973.

[99] M. E. Esandi, I. Gutierrez-Ibarluzea, N. Ibargoyen-Roteta, and B. Godman, "An evidence-based framework for identifying technologies of no or low-added value (NLVT)," *International Journal of Technology Assessment in Health Care*, vol. 36, no. 1, pp. 50–57, Jan. 2020, https://doi.org/10.1017/S0266462319000734.

[100] T. Taipalus, V. Isomottonen, H. Erkkila, and S. Ayramo, "Data Analytics in Healthcare: A Tertiary Study," *SN Computer Science*, vol. 4, no. 1, Dec. 2022, Art. no. 87, https://doi.org/10.1007/s42979-022-01507-0.

[101] K. Archer, R. Savage, S. Sanghera-Sidhu, E. Wood, A. Gottardo, and V. Chen, "Examining the effectiveness of technology use in classrooms: A tertiary meta-analysis," *Computers & Education*, vol. 78, pp. 140–149, Sep. 2014, https://doi.org/10.1016/j.compedu.2014.06.001.

[102] K. S. S. Santos, L. B. L. Pinheiro, and R. S. P. Maciel, "Interoperability Types Classifications: A Tertiary Study," in *XVII Brazilian Symposium on Information Systems*, Uberlandia, Brazil, Jun. 2021, pp. 1–8, https://doi.org/10.1145/3466933.3466952.

[103] T. Cerny, A. Al Maruf, A. Janes, and D. Taibi, "Microservice Anti-Patterns and Bad Smells. How to Classify, and How to Detect Them. A Tertiary Study." Social Science Research Network, Rochester, NY, USA, Jan. 18, 2023, https://doi.org/10.2139/ssrn.4328067.

[104] X. Liu *et al.*, "Research on Microservice Architecture: A Tertiary Study." Social Science Research Network, Rochester, NY, Aug. 30, 2022, https://doi.org/10.2139/ssrn.4204345.

[105] C. Alves, J. Cunha, and J. Araujo, "On the Pragmatics of Requirements Engineering Practices in a Startup Ecosystem," in *28th International Requirements Engineering Conference*, Zurich, Switzerland, Sep. 2020, pp. 311–321, https://doi.org/10.1109/RE48521.2020.00041.

[106] A. Yasin, R. Fatima, L. Wen, W. Afzal, M. Azhar, and R. Torkar, "On Using Grey Literature and Google Scholar in Systematic Literature Reviews in Software Engineering," *IEEE Access*, vol. 8, pp. 36226–36243, Jan. 2020, https://doi.org/10.1109/ACCESS.2020.2971712.

[107] B. Napoleao, K. Felizardo, E. Souza, and N. Vijaykumar, "Practical similarities and differences between Systematic Literature Reviews and Systematic Mappings: a tertiary study," in *29th International Conference on Software Engineering and Knowledge Engineering*, Pittsburgh, PA, USA, Jul. 2017, pp. 85–90, https://doi.org/10.18293/SEKE2017-069.

[108] K. Champion, S. Khatri, and B. M. Hill, "Qualities of Quality: A Tertiary Review of Software Quality Measurement Research." arXiv, Jul. 29, 2021, https://doi.org/10.48550/arXiv.2107.13687.

[109] Y. Zhou, H. Zhang, X. Huang, S. Yang, M. A. Babar, and H. Tang, "Quality assessment of systematic reviews in software engineering: a tertiary study," in *19th International Conference on Evaluation and Assessment in Software Engineering*, Nanjing, China, Apr. 2015, pp. 1–14, https://doi.org/10.1145/2745802.2745815.

[110] S. Jiang, K. Jakobsen, J. Bueie, J. Li, and P. H. Haro, "A Tertiary Review on Blockchain and Sustainability With Focus on Sustainable Development Goals," *IEEE Access*, vol. 10, pp. 114975–115006, Jan. 2022, https://doi.org/10.1109/ACCESS.2022.3217683.

[111] G. Lacerda, F. Petrillo, M. Pimenta, and Y. G. Gueheneuc, "Code smells and refactoring: A tertiary systematic review of challenges and observations," *Journal of Systems and Software*, vol. 167, Sep. 2020, Art. no. 110610, https://doi.org/10.1016/j.jss.2020.110610.

[112] K. Feichtinger, K. Meixner, R. Rabiser, and S. Biffl, "A Systematic Study as Foundation for a Variability Modeling Body of Knowledge," in *47th Euromicro Conference on Software Engineering and Advanced Applications*, Palermo, Italy, Sep. 2021, pp. 25–28, https://doi.org/10.1109/SEAA53835.2021.00012.

[113] M. Alghobiri, "A Comparative Analysis of Classification Algorithms on Diverse Datasets," *Engineering, Technology & Applied Science Research*, vol. 8, no. 2, pp. 2790–2795, Apr. 2018, https://doi.org/10.48084/etasr.1952.

AUTHORS PROFILE

Ghader R. Kurdi works as an assistant professor in the College of Computing at Umm Al-Qura University. Holding a Ph.D. from the University of Manchester, she was recognized with the Steve Furber Medal for her outstanding doctoral thesis. Dr. Kurdi's current research focuses on systematic reviews, specifically advancing automation methodologies for their conduct and evaluation. Beyond this, her scholarly pursuits extend to diverse topics within artificial intelligence, showcasing a broad and dynamic engagement with cutting-edge research. Her publications have received over 600 citations.

Budoor A. Allehyani works as an assistant professor in the College of Computing, Softwrare Engineering Department, at Umm Al-Qura University. She received her Ph.D. in 2018 and M.S. degree in 2012 from the University of Leicester, UK. Dr Allehyani's current research interests are in the areas of software sustainability requirements, software quality assurance, self-adaptive systems, and AI engineering.