A Systematic Literature Review on Explainability for Machine/Deep Learning-based Software Engineering Research

SICONG CAO, Yangzhou University, China
XIAOBING SUN*, Yangzhou University, China
RATNADIRA WIDYASARI, Singapore Management University, Singapore
DAVID LO, Singapore Management University, Singapore
XIAOXUE WU, Yangzhou University, China
LILI BO, Yangzhou University, China
JIALE ZHANG, Yangzhou University, China
BIN LI, Yangzhou University, China
WEI LIU, Yangzhou University, China
DI WU, University of Southern Queensland, Australia
YIXIN CHEN, Washington University in St. Louis, USA

The remarkable achievements of Artificial Intelligence (AI) algorithms, particularly in Machine Learning (ML) and Deep Learning (DL), have fueled their extensive deployment across multiple sectors, including Software Engineering (SE). However, due to their black-box nature, these promising AI-driven SE models are still far from being deployed in practice. This lack of explainability poses unwanted risks for their applications in critical tasks, such as vulnerability detection, where decision-making transparency is of paramount importance. This paper endeavors to elucidate this interdisciplinary domain by presenting a systematic literature review of approaches that aim to improve the explainability of AI models within the context of SE. The review canvasses work appearing in the most prominent SE & AI conferences and journals, and spans 63 papers across 21 unique SE tasks. Based on three key Research Questions (RQs), we aim to (1) summarize the SE tasks where XAI techniques have shown success to date; (2) classify and analyze different XAI techniques; and (3) investigate existing evaluation approaches. Based on our findings, we identified a set of challenges remaining to be addressed in existing studies, together with a roadmap highlighting potential opportunities we deemed appropriate and important for future work.

*Corresponding author

Authors' addresses: Sicong Cao, DX120210088@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Xiaobing Sun, xbsun@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Ratnadira Widyasari, ratnadiraw.2020@phdcs.smu.edu.sg, School of Computing and Information Systems, Singapore Management University, Singapore; David Lo, davidlo@smu.edu.sg, School of Computing and Information Systems, Singapore Management University, Singapore; Xiaoxue Wu, xiaoxuewu@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Lili Bo, lilibo@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Jiale Zhang, jialezhang@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Bin Li, lb@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Wei Liu, weiliu@yzu.edu.cn, School of Information Engineering, Yangzhou University, Yangzhou, China; Di Wu, School of Mathematics, Physics, and Computing, University of Southern Queensland, Toowoomba, Australia, di.wu@unisq.edu.au; Yixin Chen, Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, USA, chen@cse.wustl.edu.

Manuscript submitted to ACM

CCS Concepts: • **General and reference** → *Surveys and overviews*; • **Computing methodologies** → *Neural networks*; *Artificial intelligence*; • **Software and its engineering** → *Software development techniques.*

Additional Key Words and Phrases: Explainable AI, XAI, interpretability, neural networks, survey

ACM Reference Format:
Sicong Cao, Xiaobing Sun, Ratnadira Widyasari, David Lo, Xiaoxue Wu, Lili Bo, Jiale Zhang, Bin Li, Wei Liu, Di Wu, and Yixin Chen. 2023. A Systematic Literature Review on Explainability for Machine/Deep Learning-based Software Engineering Research. 1, 1 (December 2023), 7 pages. https://doi.org/XXXXXXX.XXXXXXX

A REVIEW METHODOLOGY

A.1 Search Strategy

As shown in Fig. 1, following the standardized practice within the field of SE, our first step involves identifying primary studies to enhance our ability to address the formulated RQs effectively. Given that the DL revolution – triggered by AlexNet in 2012 – has transformed AI research and became the catalyst for the ML/DL boom in all fields including SE, we chose a 13-year period of January 1st, 2012, to December 31st, 2024, to collect the literature related to XAI4SE. Next, we identified the top peer-review and influential conference and journal venues in the domains of SE and Programming Languages (PL), as outlined in Table 1. In total, we included 16 conferences (ICSE, ASE, ESEC/FSE, ICSME, ICPC, RE, ESEM, ISSTA, MSR, SANER, ISSRE, APSEC, COMPSAC, QRS, OOPSLA, PLDI) and six journals (TSE, TOSEM, EMSE, JSS, IST, ASEJ). We chose to include PL venues in our study given the frequent overlap of SE and PL research. Furthermore, we also include five top conferences (AAAI, ICML, ICLR, NeurIPS, IJCAI) that centered on machine learning (ML) and deep learning (DL) as these conferences might feature papers applying ML and DL techniques to SE tasks.

Apart from manually searching primary studies from top-tier venues, we also retrieved relevant papers from five popular digital libraries, including IEEE Xplore[1], ACM Digital Library[2], SpringerLink[3], Wiely[4], and Scopus[5], and two of the most popular research citation engines, Web of Science[6] and Google Scholar[7], based on the search string (listed in Table 2) assembled from a group of topic-related keywords summarized from manually collected papers. As shown in Table 3, we collected a total of 70,325 relevant studies with the automatic search from these seven electronic databases.

A.2 Study Selection

A.2.1 Inclusion and Exclusion Criteria. After paper collection, we performed a relevance assessment according to the following inclusion and exclusion criteria:

- ✓ *The paper must be written in English.*
- ✓ *The paper must be a peer-reviewed full research paper published in a conference proceeding or a journal.*
- ✓ *The paper must have an accessible full text.*
- ✓ *The paper must adopt ML/DL techniques to address SE problems.*
- ✗ *The paper has less than 6 pages.*
- ✗ *Books, keynote records, non-published manuscripts, and grey literature are dropped.*

[1] https://ieeexplore.ieee.org
[2] https://dl.acm.org
[3] https://link.springer.com
[4] https://onlinelibrary.wiley.com
[5] https://www.scopus.com
[6] https://www.webofscience.com
[7] https://scholar.google.com

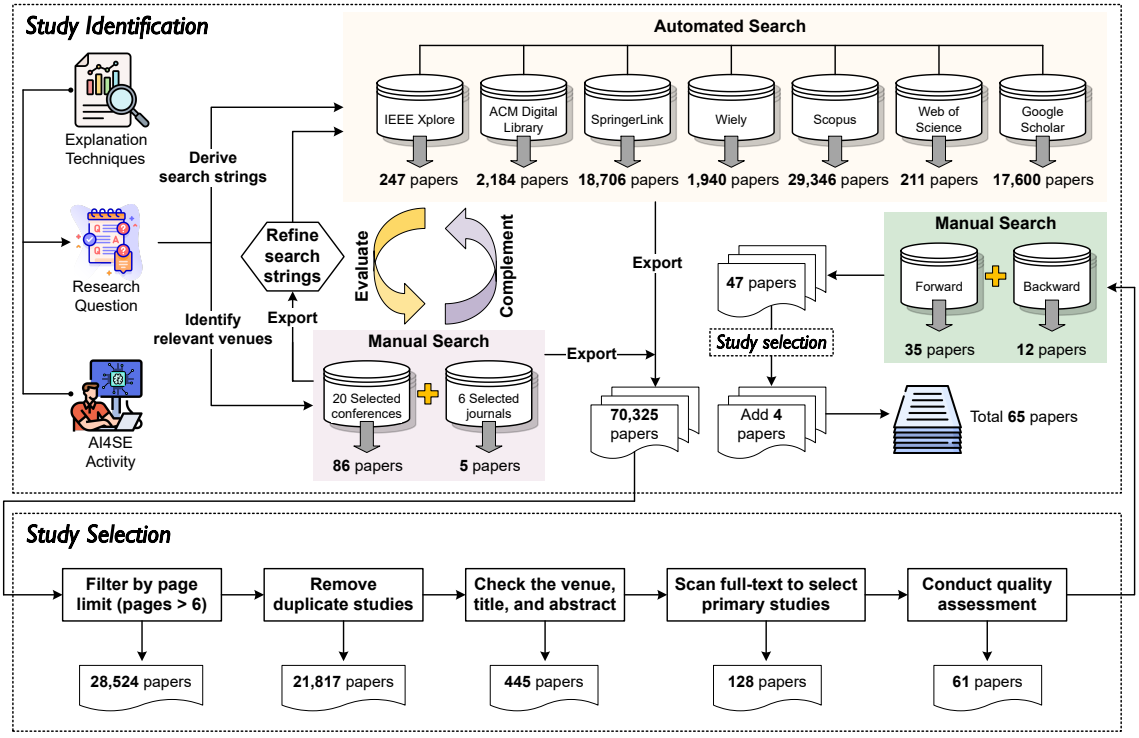


Fig. 1. Study identification and selection process.

- ✗ *The paper is a literature review or survey.*
- ✗ *The paper is not a conference paper that has been extended as a journal paper.*
- ✗ *The paper uses SE approaches to contribute to ML/DL systems.*
- ✗ *The studies that do not apply XAI techniques on SE tasks are ruled out.*
- ✗ *The studies where explainability is discussed as an idea or part of the future work are excluded.*

In particular, by literature filtering and deduplication (exclusion criteria 1), the total number of included papers was reduced to 21,817. After the first two authors manually examined the venue, title, and abstracts of the papers, the total number of included papers declined substantially to 445. Any ambiguous papers would be forwarded to the fourth and 11th authors who were experienced in the fields of SE and XAI research to conduct a secondary review. In addition, books, keynote records, non-published manuscripts, grey literature, SLRs/surveys, and conference versions of extended papers were also discarded in this phase (exclusion criteria 2-4). The SE4AI papers [1], which used SE approaches to contribute to ML/DL systems were also not considered (exclusion criteria 5) because our SLR focused exclusively on the explainability of AI4SE models. Furthermore, we ruled out studies that did not apply XAI techniques on SE tasks, or just discussed explainability as an idea or future work (exclusion criteria 6,7). In the fourth phase, we reviewed the full texts of the papers (inclusion criteria 3), identifying 128 primary studies directly relevant to our research topic.

A.2.2 Quality Assessment. To prevent biases introduced by low-quality studies, we formulated five **Quality Assessment Criteria (QAC)**, given in Table 4, to evaluate the 128 included studies. The quality assessment process was

Table 1. Publication Venues for Manual Search

Venue	Acronym	Full name
Conference	ICSE	IEEE/ACM International Conference on Software Engineering
	ASE	IEEE/ACM International Conference Automated Software Engineering
	ESEC/FSE	ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering
	ICSME	IEEE International Conference on Software Maintenance and Evolution
	ICPC	IEEE International Conference on Program Comprehension
	RE	IEEE International Conference on Requirements Engineering
	ESEM	ACM/IEEE International Symposium on Empirical Software Engineering and Measurement
	ISSTA	ACM SIGSOFT International Symposium on Software Testing and Analysis
	MSR	IEEE Working Conference on Mining Software Repositories
	SANER	IEEE International Conference on Software Analysis, Evolution and Reengineering
	ISSRE	IEEE International Symposium on Software Reliability
	APSEC	Asia-Pacific Software Engineering Conference
	COMPSAC	IEEE International Computer Software and Applications Conference
	QRS	IEEE International Conference on Software Quality, Reliability and Security
	OOPSLA	ACM SIGPLAN International Conference on Object-oriented Programming, Systems, Languages, and Applications
	PLDI	ACM SIGPLAN Conference on Programming Language Design and Implementation
	AAAI	AAAI Conference on Artificial Intelligence
	ICML	International Conference on Machine Learning
	ICLR	International Conference on Learning Representations
	NeurIPS	Annual Conference on Neural Information Processing Systems
	IJCAI	International Joint Conference on Artificial Intelligence
Journal	TSE	IEEE Transactions on Software Engineering
	TOSEM	ACM Transactions on Software Engineering and Methodology
	EMSE	Empirical Software Engineering
	JSS	Journal of Systems and Software
	IST	Information and Software Technology
	ASEJ	Automated Software Engineering

Table 2. Search Keywords

Group	Keywords
1	*"Machine Learn*"* OR *"Deep Learning"* OR *"Neural Network?"* OR *"Reinforcement Learning"*
2	*"Explainable"* OR *"Interpretable"* OR *"Explainability"* OR *"Interpretability"*
3	*"Software Engineering"* OR *"Software Analytics"* OR *"Software Mainten*"* OR *"Software Evolution"* OR *"Software Test*"* OR *"Software Requirement?"* OR *"Software Develop*"* OR *"Project Management"* OR *"Software Design*"* OR *"Dependability"* OR *"Security"* OR *"Reliability"*
4	*"Code Representation"* OR *"Code Generation"* OR *"Code Comment Generation"* OR *"Code Search"* OR *"Code Localization"* OR *"Code Completion"* OR *"Code Summarization"* OR *"Method Name Generation"* OR *"Bug"* OR *"Fault"* OR *"Vulnerability"* OR *"Defect"* OR *"Test Case"* OR *"Program Analysis"* OR *"Program Repair"* OR *"Clone Detection"* OR *"Code Smell"* OR *"SATD Detection"* OR *"Compile"* OR *"Code Review"* OR *"Code Classification"* OR *"Code Change"* OR *"Incident Detection"* OR *"Effort Cost Prediction"* OR *"GitHub"* OR *"StackOverflow"* OR *"Developer"*

* is a wildcard used to match zero or more characters.
? is another wildcard used to match a single character.

piloted by the first and second authors, involving 30 randomly selected primary studies. We adopted pairwise inter-rater reliability with Cohen's Kappa statistic to measure the consistency of the markings. For any case that they did not reach

Table 3. Summary of the Process of Study Search and Selection

Data Source	# Studies
IEEE Xplore	247
ACM Digital Library	2,184
SpringerLink	18,706
Wiely	1,940
Scopus	29,346
Web of Science	211
Google Scholar	17,600
Merge	70,325
Filtering studies less than 6 pages	28,524
Removing duplicated studies	21,817
Excluding primary studies based on venue, title, and abstract	445
Excluding primary studies based on full text	128
After Quality Assessment	61
After Forward & Backward Snowballing	108
Final	63

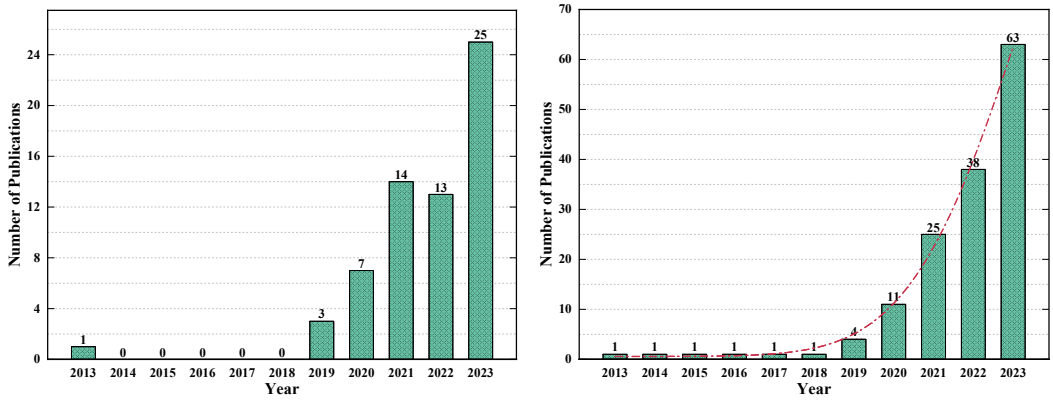
Table 4. Checklist of Quality Assessment Criteria for Explainability Studies in AI4SE

No.	Quality Assessment Criteria
QAC_1	Is the impact of the proposed approach (or empirical/case study) on the AI4SE community clearly stated?
QAC_2	Are the contributions of the study clearly claimed?
QAC_3	Does the study provide a clear description of the workflow and implementation of the proposed approach?
QAC_4	Are the experiment details, including datasets, baselines, and evaluation metrics, clearly described?
QAC_5	Do the findings drawn from the experiments strongly substantiate the arguments presented in the study?

Table 5. Extracted Data Items and Related Research Questions

RQ	Data Item
RQ_1	The SE task that an XAI4SE approach tries to solve
RQ_1	The SE activity in which each SE task belongs
RQ_1	Publication type of each primary study (i.e., new technique, empirical study, or case study)
RQ_2	XAI technique employed by each study
RQ_2	Explanation format
RQ_3	The adopted baseline approaches
RQ_3	Benchmark dataset name
RQ_3	Presence/absence of replication package
RQ_3	What metrics are used to evaluate the XAI techniques

a consensus after open discussions, the fourth and 11th authors (domain experts experienced in SE and XAI) were consulted as tie-breakers. Within two iterations, the Cohen's Kappa coefficient was successfully raised from *moderate* (0.58) to *almost perfect agreement* (0.84). Then, an assessment was performed for the remaining 98 primary studies. After quality assessment, a final set of 61 high-quality papers was reserved.



(a) Distribution of publications per year.

(b) Cumulative publication trend over years.

Fig. 2. Publication statistics over years.

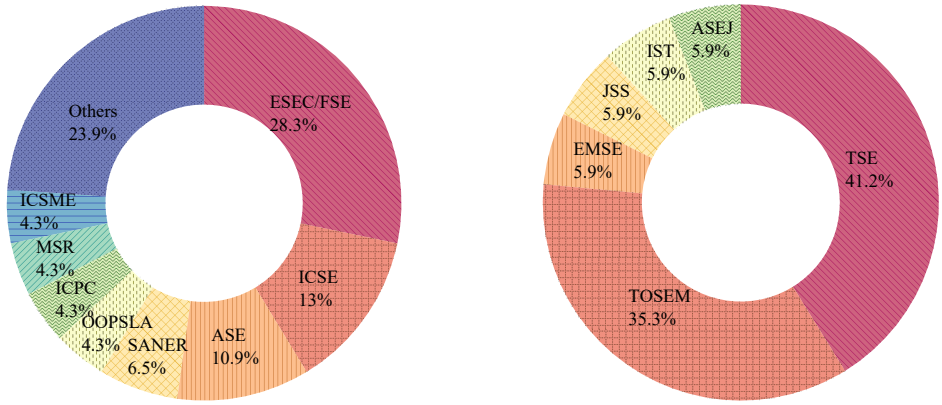
A.2.3 Forward and Backward Snowballing. To avoid omitting any possibly relevant work during our manual and automated search process, we also performed lightweight backward and forward snowballing, i.e., basically examining the research referenced in each of our selected primary studies, as well as the publications that subsequently referred to these studies, on the references and the citations of 61 high-quality papers. As a supplement, we gathered 47 more papers, and conducted the complete study selection process again, including filtering, deduplication, and quality assessment, and obtained two additional papers.

A.3 Data Extraction and Analysis

Based on the collected 63 primary studies, we extracted the essential data items used to answer three main RQs. In Table 5, we outline the details information extracted and gathered from 63 primary studies. The column labeled "Data Item" enumerates the relevant data items extracted from each primary study, while the column "RQ" specifies the corresponding research question. In order to mitigate errors during data extraction, the first two authors working together on extracting these data items from the primary studies. Then, the fifth author verified the extracted data results.

Fig. 2a presents the distribution of selected primary studies in each year. The first XAI4SE study we found was published in 2013. After that, there was a 5-year research gap, ranging from 2014 to 2018. The enthusiasm for investigating the explainability on AI4SE models has steadily risen since 2019, and reaches its peak in 2023, comprising 39.7% of the total publications. Fig. 2b illustrates the cumulative publication trend over years. It is observable that the slope of the curve fitting the distribution experiences a significant increase between 2019 and 2023. This pronounced upward trend indicates a burgeoning research interest in the field of XAI4SE.

We also analyzed the publication trend of primary studies in selected conferences and journal venues, respectively. As shown in Fig. 3a, ESEC/FSE stands out as the predominant conference venues favored by XAI4SE studies, with a contribution of 28.3% of the total. Other venues making noteworthy contributions include ICSE (13%), ASE (10.9%), and SANER (6.5%). Fig. 3b shows the distribution of primary papers published in different journal venues. It can be seen that 76.5% of relevant papers were published in TSE and TOSEM, which indicates a booming trend of XAI4SE research in top-tier SE journals in the past few years.



(a) Distribution of publications in various conferences.

(b) Distribution of publications in various journals.

Fig. 3. Distribution of selected papers in different publication venues.

REFERENCES

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald C. Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: a case study. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE/ACM, 291–300.