Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

A CONSISTENCY-AWARE PERSPECTIVE ON EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR FEATURE SELECTION IN SOFTWARE ENGINEERING: A CRITICAL REVIEW AND FRAMEWORK

[1]Adam Khan, [2]Asad Ali, [3]Muhammad Ismail Mohmand

[1]*Department of Computer Science, Sarhad University of Science and Information Technology Peshawar, Pakistan*

[2]*Computer Engineering Department, Cyprus International University, Nicosia, North Cyprus*

[3]*Department of Computer Engineering at the Faculty of Engineering,and Natural Sciences at Istanbul Atlas University, 34408, Turkey.*

[1]adam.me@suit.edu.pk, [2]aali@ciu.edu.tr, [3]muhammad.mohmand@atlas.edu.tr

Keywords
Explainable Artificial Intelligence (XAI), Software Engineering, Feature Selection Consistency, Model-Agnostic Explainability, Permutation Feature Importance (PFI)

Article History
Received on 14 Feb, 2026
Accepted on 16 March, 2026
Published on 18 March, 2026

Corresponding Author:
Adam Khan

Abstract

Explainable Artificial Intelligence (XAI) is become essential to enhance transparency, interpretability and trust in Machine Learning (ML) models in Software Engineering (SE). Although model-agnostic approaches such as Local Interpretable Model-Agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP) and Permutation Feature Importance (PFI) are increasingly popular for prediction interpretation, their effectiveness in assessing Feature Selection (FS) is an issue of serious concern. Specifically, the ranking of feature importance produced by these methods is often unstable across changes in datasets, model configurations, and validation techniques, and has less practical application in SE decision-making. This study presents a critical and thematic review of XAI methods for FS in SE , with particular emphasis on the explanation consistency. Unlike prior studies, it methodologically examines the shortcomings of current methods in terms of consistency. On the basis of the identified research gaps, we propose the CFXAI-SE framework (Consistent Feature eXplainable AI for Software Engineering). The framework combines dataset perturbation, multi-model analysis and statistical consistency analysis to produce a consistent and reliable feature importance ranking. The findings reveal that consistency is largely unexplored aspect in XAI studies for SE. The proposed framework suggested a systematic context in building reliable, interpretable, and reproducible ML systems. This study contributes to advancing dependable XAI implementation in SE applications, such as defect prediction and effort estimation.

1. Introduction

Machine Learning (ML) has been incorporated into Software Engineering (SE), and has significantly advanced predictive capabilities, enabling defect prediction, fault localization, effort estimation, and software quality assessment [1][2]. Regardless of these improvements, many high performing ML models, particularly ensemble methods and deep learning models, are "*black boxed*", offering little insight into how they make decisions internally [3]. Such a lack of transparency poses significant risks of trust, accountability and adoption, particularly within industrial SE settings.

Explainable Artificial Intelligence (XAI) has become a promising paradigm that can be utilized to resolve these challenges by making model behavior more comprehensible to humans [4]. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) [5], SHapley Additive exPlanations (SHAP) [6], Permutation Feature Importance (PFI) [7], are capable of providing insight into predictions and the relative feature contributions. Such techniques are increasingly adopted in SE to assist in debugging, model validation and informed decision making.

Nevertheless, the consistency of Feature Selection (FS) explanations remains a critical challenge that has been addressed insufficiently. In practice, feature ranking results produced by XAI techniques can vary significantly with even minor alterations in training data, model parameters, or validation techniques. This instability makes the explanations unreliable and may cause inconsistent or misleading conclusions in SE applications [8].

Despite the fact that previous studies have investigated the role of XAI in SE, most have concentrated on interpretability and predictive performance, while ignoring reproducibility and consistency of explanations. Addressing this gap requires the development of Consistency-aware XAI methodologies tailored to SE.

This study makes the following key contributions:

- C1: Provides thematic and critical review of XAI techniques for FS in SE.
- C2: Identifies and examines the issue of inconsistency in features importance explanations.
- C3: Generalizes the discussion across domains to contextualize challenges of XAI in SE.
- C4: Introduces the CFXAI-SE framework, a new methodology towards a stable and reproducible FS outcomes.

2. Thematic Literature Review

2.1 Overview of XAI in Software Engineering

Explainable Artificial Intelligence (XAI) has attracted considerable attention as a solution to the persistent problem of opacity in machine learning (ML) models [9][10]. Within SE, XAI is becoming widely applied to various tasks such as defect prediction, fault localisation, effort estimation and software analytics [11]. XAI techniques will address this issue by offering interpretable model behaviour information to achieve the technical and managerial decision-making of SE contexts.

The growing significance of XAI is also evidenced by the fact that its study and public attention are increasing over the years. Both scholarly articles and online search behaviors show that interest in XAI is growing dramatically after 2019, as seen in Figure 1. This trend is closely linked to the rise of ethical AI awareness, regulatory demands, or the necessity of ML systems transparency [12].

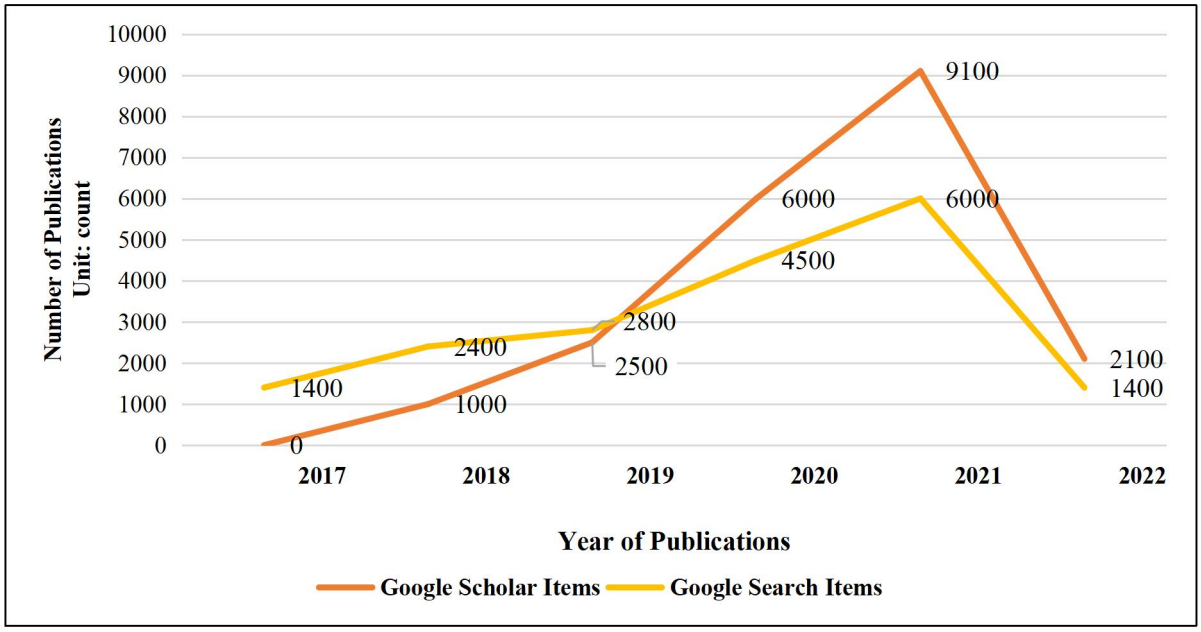


Figure 1. Popularity index of XAI measured through Google Scholar publications and Google search trends (2017–2022) [12]

As illustrated in Figure 2, there is a notable surge in the number of publications associated with XAI across several key academic databases. That is, between 2014 and 2024, IEEE Xplore, PubMed, and Nature Portfolio will witness substantial growth in XAI-related publications. For instance, The findings show that there has been a gradual increase in XAI-related research in domains as shown in Figure 2 PubMed has published 916 items (primarily reviews and assessment studies) between 2014 and 2024, IEEE Xplore 3,824 items (with a high volume of technical application), and Nature Portfolio 216 high-impact items (some of them reviews and original work). In short, the increase in the number of publications in IEEE Xplore, increased from zero in 2014 to 1,319 in 2024. This observable increase brings into focus the fast growing use of XAI and its growing capabilities to deal with complex real-world tasks particularly in SE, that include defect predictions, fault localization, and efforts estimation. This consistent rise demonstrates the significance of XAI in both engineering and interdisciplinary disciplines. Further, it reflects the continuing importance of XAI in SE research.

In addition, according to existing studies, interpretability is crucial in enhancing trust, debugging, and decision-making [13], [14]. Nevertheless, the implementation of XAI in SE is still in its initial phase as compared to other domains such as healthcare and finance [15].

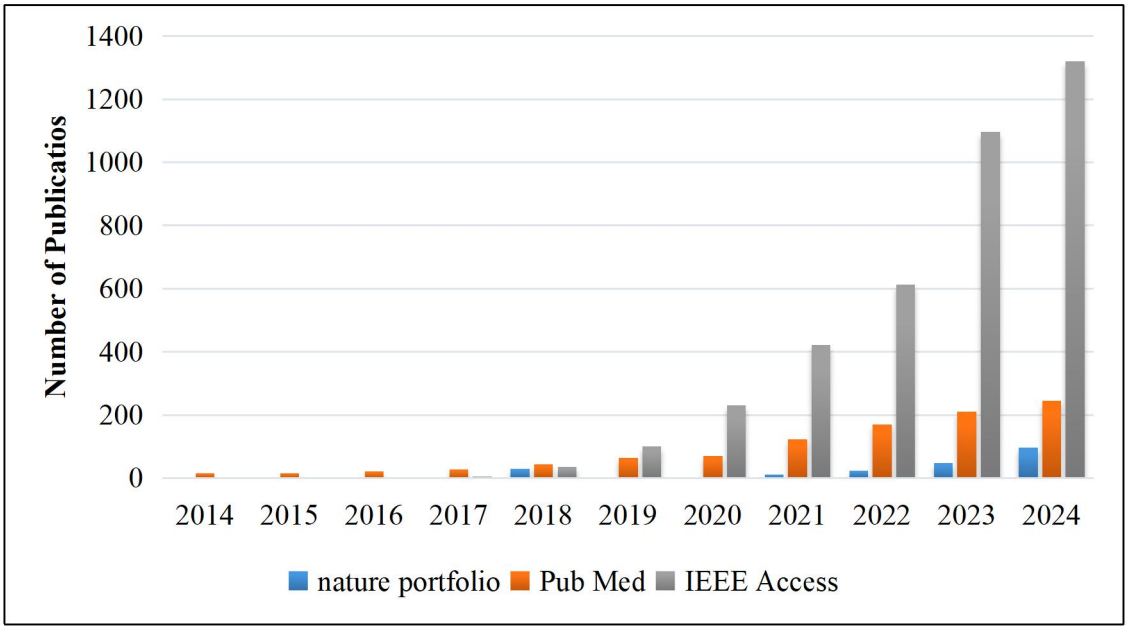


Figure 2 Number of publications (annual) containing the term "Explainable Artificial Intelligence" across IEEE Xplore, PubMed, and Nature Portfolio databases (2014-2024)

2.2 Model-Agnostic XAI Techniques for Feature Selection

The popularity of model-agnostic XAI techniques is because they can be applied to a variety of ML models. Few widely adopted methods are given below:

- LIME (Local Interpretable Model-Agnostic Explanations) generates local surrogate models to explain individual predictions. Although efficient, its explanations can be unstable due to sensitivity to random sampling and perturbations [16].
- SHAP (SHapley Additive exPlanations) is a theoretically grounded feature attribution method based on cooperative game theory. It ensures local accuracy of explanations, but its computational cost is high and results are sensitive to the choice of background data [17].
- Permutation Feature Importance (PFI) compares the importance of features by comparing the performance loss during shuffling of features. Even though it is intuitive, it is influenced by feature correlation and dataset characteristics [7].
- BreakDown and iBreakDown, methods offer sequential and interactive explanations, which increase the interpretability of end-users but are not as robust as validation techniques [18][19].

These techniques are commonly used in SE, but they are mainly concerned with interpretability, not consistency. In order to summarize the main features of commonly used XAI methods in SE, Table 1 provides a comparative overview of their features and applicability to SE tasks.

Table 1: ***XAI Methods Used in Software Engineering***

Methodology	Key Features	Relevance to SE
LIME	Local, interpretable explanations	Enhances decision transparency
SHAP	Feature attribution using game theory	Provides consistent explanations
LIME HPO	Hyperparameter optimization	Improves explanation stability
BreakDown	Sequential feature contribution	Useful for fault prediction
iBreakDown	Interactive explanations	Supports stakeholder understanding

2.3 Role of Feature Selection in Software Engineering

In SE tasks, FS is an important component of enhancing model performance and model interpretability [20][21]. It assists in finding important software metrics including:

- Code complexity
- Change frequency
- Developer activity

The features are necessary in order to predict defects, estimate effort, and measure software quality [22].

Traditional FS (e.g., correlation-based, wrapper methods) methods are performance optimizing, however, they have no interpretability. XAI-based FS approaches offer comprehensible insights, which is more appropriate in the context of SE decision-making. Nevertheless, the unstable FS results may lead to false conclusions, particularly when the rankings of features are not consistent in various experimental circumstances [8].

2.4 Consistency and Stability Issues in XAI

One of the critical drawbacks of current XAI methods is the absence of stability in the ranking of the importance of the features. Research demonstrates that the explanations may vary because of:

- Data perturbations (sampling, noise, imbalance)
- Model variability (algorithms, hyperparameters)
- Random initialization and training processes [8][23]

This inconsistency leads to the question of the reliability and reproducibility of XAI outputs, especially in SE, where decisions are made based on consistency feature importance [8][24].

2.5 Evaluation Metrics and Limitations

In the majority of XAI studies, models are assessed in terms of:

- Interpretability
- Fidelity
- Computational efficiency

Nevertheless, consistency measures are not often taken into consideration [25].

Commonly used measures of stability are:

- **Spearman Rank Correlation** – measures ranking similarity
- **Kendall's Tau** – evaluates ordinal association
- **Jaccard Similarity** – compares feature subsets

These measures are not widely incorporated systematically in XAI processes, which is a significant gap in research [26][27].

2.6 Cross-Domain Insights

SE can learn important lessons through other domains:

- **Healthcare**: Requires highly reliable explanations for clinical decisions [28]
- **Finance**: Emphasizes regulatory compliance and transparency [29]
- **Cybersecurity**: Focuses on interpretability for threat analysis [30]

The focus of these areas emphasizes the need to have strong, stable, and reliable explanations, which are not yet present in SE applications.

Better insight into the wider issues of XAI and the effect of the concept on SE can be provided by examining how explainability is considered in various fields. Table 2 provides a comparative overview of the key issues in those areas and explains their applicability to SE. This comparison shows that factors like interpretability, fairness, computational efficiency and reliability are not domain specific, but are inherent problems of XAI systems.

Although XAI is increasingly applied in other domains, there are still a number of constraints to applying it to SE. The current literature mainly addresses the interpretability and predictive performance, whereas the consistency of FS explanations is largely underexploited. Even though there are statistical indicators of stability, they are poorly systematically incorporated into XAI workflows. Moreover, findings in other domains indicate the existence of common issues, but their implications to FS consistency in SE are insufficiently formulated. All these constraints negatively affect the practicability of XAI applications in SE and encourage the development of a consistency-aware model, which is suggested in the present study.

Table 2: ***Comparative Summary of XAI Challenges Across Domains***

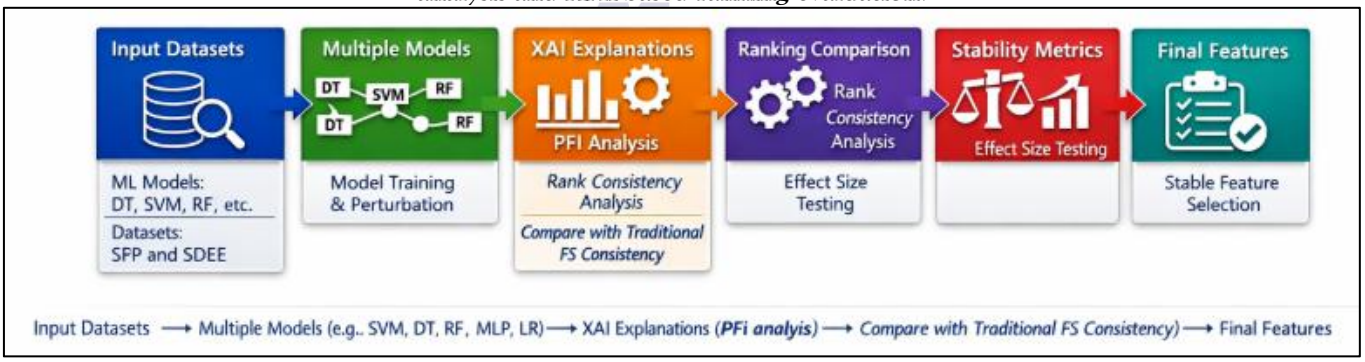
Domain	Main XAI Challenges	Similar Gap in SE?	Notes / Relevance to SE
Transportation	Real-time interpretability, computational efficiency	Yes	SE requires interpretable models for real-time fault detection
Healthcare	Stakeholder-specific	Yes	SE needs clear explanations for

	explanations, privacy, regulatory compliance		developers and users
Law	Fairness, ethical compliance, interpretive consistency	Yes	SE must ensure fairness in decision-support systems
Finance	Dynamic decisions, accountability, complexity	Yes	SE requires efficient and interpretable models
Cybersecurity	False positives/negatives, lack of standards	Yes	Similar issues in vulnerability detection
Software Engineering	**Feature selection consistency**	**N/A**	**Focus of this study**

3. Proposed Framework: CFXAI-SE

The presented CFXAI-SE framework is aimed at achieving consistent FS by introducing consistency as one of the fundamental evaluation measures. The framework starts with the input SE datasets and then proceeds with preprocessing and dataset perturbation, in order to have real-world variability, as shown in Figure 3. Various ML models (e.g., Decision Tree, SVM, Random Forest) are then trained in order to have some model diversity. The model-agnostic XAI, particularly PFI, is used to obtain the rankings of feature importance. These rankings are then compared with evaluation of consistency such as rank comparison and statistical stability measures like effect size tests. The framework also provides the ability to compare it with the traditional methods of FS to determine robustness. Integrating perturbation, multi-model analysis, and statistical validation, CFXAI-SE overcomes the major weaknesses of the current XAI methods and improves the consistency of FS in SE work.

Figure 3. Workflow of the CFXAI-SE framework for consistency-aware feature selection using multi-model analysis and XAI-based ranking evaluation.



4. Conclusion and Future Work

In this study, the consistent critical review of XAI techniques in SE has been outlined where the present methods yield interpretability, but lack consistency. The study revealed the key limitations, such as the lack of validation on multi-models, insufficient application of statistical stability measures, and non-coherent framework, through a thematic analysis and cross-domain understanding. To overcome these issues, it was suggested to use the CFXAI-SE framework that combines the perturbation of datasets, multi-model learning, and consistency checking to produce dependable and understandable feature importance rankings. The framework increases credibility and assists in making more informed decisions on SE tasks such as defect prediction and effort estimation. Future directions include empirical validation of the framework on a wider range of datasets and expansion of the framework to complex models, which will further justify the use of robust and reliable XAI in SE domain.

References

[1] Wang, S., Huang, L., Gao, A., Ge, J., Zhang, T., Feng, H., ... & Ng, V. (2022). Machine/deep learning for software engineering: A systematic literature review. *IEEE Transactions on Software Engineering*, *49*(3), 1188-1231.

[2] Sakhrawi, Z., Sellami, A., & Bouassida, N. (2021). Software enhancement effort prediction using

machine-learning techniques: A systematic mapping study. *SN Computer Science*, *2*(6), 468.

[3] Apu, K. U., Rahman, M. M., Hoque, A. B., & Bhuiyan, M. (2022). Forecasting future investment value with machine learning, neural networks, and ensemble learning: a meta-analytic study. *Review of Applied Science and Technology*, *1*(02), 01-25.

[4] Khan, A., Ali, A., Khan, J., Ullah, F., & Faheem, M. (2025). Using Permutation-Based feature importance for improved machine learning model performance at reduced costs. *IEEE Access*.

[5] Jiarpakdee, J., Tantithamthavorn, C. K., Dam, H. K., & Grundy, J. (2020). An empirical study of model-agnostic techniques for defect prediction models. *IEEE Transactions on Software Engineering*, *48*(1), 166-185.

[6] Kassaymeh, S., Rjoub, G., Dssouli, R., Bentahar, J., & Almobydeen, S. B. (2024, August). Interpretable shap-driven machine learning for accurate fault detection in software engineering. In *Joint International Conference on AI, Big Data and Blockchain* (pp. 52-66). Cham: Springer Nature Switzerland.

[7] Abdelaziz, M. T., Radwan, A., Mamdouh, H., Saad, A. S., Abuzaid, A. S., AbdElhakeem, A. A., ... & Darweesh, M. S. (2025). Enhancing network threat detection with random forest-based NIDS and permutation feature importance. *Journal of Network and Systems Management*, *33*(1), 2.

[8] Khan, A., Ali, A., Khan, J., Ullah, F., & Faheem, M. (2025). Exploring Consistent Feature Selection for Software Fault Prediction: An XAI-based model-agnostic Approach. *IEEE Access*.

[9] Thalpage, N. (2023). Unlocking the black box: Explainable artificial intelligence (XAI) for trust and transparency in ai systems. *J. Digit. Art Humanit*, *4*(1), 31-36.

[10] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503-3568.

[11] Cao, S., Sun, X., Widyasari, R., Lo, D., Wu, X., Bo, L., ... & Chen, Y. (2025). A Systematic Literature Review on Explainability for ML/DL-based Software Engineering. *ACM Computing Surveys*, *58*(4), 1-34.

[12] S. M. Hussain *et al.*, "Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images: The Role of Explainable Artificial Intelligence," *Applied Sciences*, vol. 12, no. 12, Art. no. 12, Jan. 2022, doi: 10.3390/app12126230.

[13] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, *64*(12), 3197-3234.

[14] Barnes, E., & Hutson, J. (2024). Navigating the complexities of AI: The critical role of interpretability and explainability in ensuring transparency and trust. *International Journal of Multidisciplinary and Current Educational Research*, *6*(3).

[15] Gupta, N. (2025). Explainable AI for regulatory compliance in financial and healthcare sectors: a comprehensive review. *International Journal of Advances in Engineering and Management*, *7*(3), 489-494.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Aug. 09, 2016, *arXiv*: arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938.

[17] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: June 29, 2025. [Online].

[18] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022).
Explainable AI methods: A brief overview. In *xxAI – Beyond explainable AI* (pp. 13–38). Springer.
https://doi.org/10.1007/978-3-031-04083-2_2

[19] Gosiewska, A., Biecek, P.: iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models. arXiv preprint arXiv:1903.11420 (2019)

[20] Balogun, A. O., Basri, S., Mahamad, S., Abdulkadir, S. J., Almomani, M. A., Adeyemo, V. E., ... & Bajeh, A. O. (2020). Impact of feature selection methods on the predictive performance

of software defect prediction models: an extensive empirical study. *Symmetry*, *12*(7), 1147.

[21] Ali, M., Mazhar, T., Shahzad, T., Ghadi, Y. Y., Mohsin, S. M., Akber, S. M. A., & Ali, M. (2023). Analysis of feature selection methods in software defect prediction models. *IEEE Access*, *11*, 145954-145974.

[22]Gao, K., Khoshgoftaar, T. M., Wang, H., & Seliya, N. (2011). Choosing software metrics for defect prediction: an investigation on feature selection techniques. *Software: Practice and Experience*, *41*(5), 579-606.

[23] Ribeiro, J., Cardoso, L., Santos, V., Carvalho, E., Carneiro, N., & Alves, R. (2024). How Reliable and Stable are Explanations of XAI Methods?. *arXiv preprint arXiv:2407.03108*.

[24] ŞAHiN, E., Arslan, N. N., & Özdemir, D. (2025). Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural computing and applications*, *37*(2), 859-965.

[25] Kadir, M. A., Mosavi, A., & Sonntag, D. (2023, July). Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)* (pp. 000111-000124). IEEE.

[26] Pawlicki, M. (2023, October). Towards quality measures for xAI algorithms: Explanation stability. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1-10). IEEE.

[27] Balestra, C. (2024). *Rankings and importance scores as multi-facets of explainable machine learning* (Doctoral dissertation, Dissertation, Dortmund, Technische Universität, 2024).

[28] Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics* (pp. 160-169). IEEE.

[29] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, *57*(8), 216.

[30] Almheiri, S. J., Shah, A. A., Abbas, S., Ahmad, M., & Khan, M. A. (2025). Smart sustainable cyber security: modelling an interpretable and transparent threat detection with explainable artificial intelligence. *Discover Sustainability*, *6*(1), 442.