***XAI in SLR***

***Report 1***

1. **“Systematic Literature Review of Explainable AI for Software Engineering”**

**Scope of the study:**An analysis of the applications of XAI in Software Engineering (SE). The study reviewed 24 papers identified from an initial pool of 869 primary studies.

**Key findings:**The most frequently studied area of SE was software maintenance, particularly defect prediction (68% of the papers).  
XAI was primarily applied in classical ML models, such as logistic regression, with limited use of advanced DL models.  
A lack of unified evaluation metrics for XAI in the context of SE was highlighted.

**Identified challenges:**Low level of XAI applications in tasks such as software testing and debugging.  
Scarcity of explanation methods presented in visual formats or natural language.  
Limited standards for evaluating the usability of XAI.

**Recommendations:**Development of more universal and comprehensible explanation methods.  
Further research into the usability of XAI in SE tasks, such as bug fixing and testing.

1. <https://www.mdpi.com/2076-3417/14/19/8884>

**Recent Applications of Explainable AI (XAI): A Systematic Literature Review**

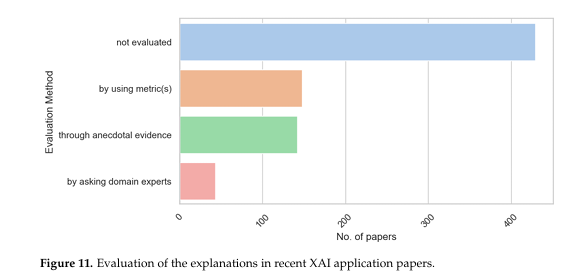
Need for XAI – importance of providing transparency and interpretability to AI models, because of their fields of applications – Health-related (oncology, diagnostics) and other high-stakes areas.

Key application areas of XAI (in chosen surveys):

* **Healthcare**: Cancer diagnosis, COVID-19 detection, diagnostic medical imaging.
* **Agriculture and Environment**: Water quality monitoring, plant disease identification.
* **Cybersecurity**: Intrusion detection systems, source code vulnerability analysis.
* **Finance**: Credit risk assessment, fraud detection.
* **Education and Law**: Academic success prediction, judicial decision analysis.

SHAP vs. LIME

|  |  |
| --- | --- |
| **SHapley Additive exPlanations** | **Local Interpretable Model-agnostic Ex planations** |
| Both model-agnostic XAI tools & model-specific XAI tools | Fully model-agnostic (independent of the prediction model and can be used on top of any linear or non-linear model |
| Efficiency, symmetry, dummy, additivity | Criticized for its instability (same inputs != same outputs |
| Mathematical guarantees to address local-to-global limitation | Local approximation lacks stable connection to the global level of the model |



The most commonly used XAI techniques:

* SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were the most popular explanation methods.
* Gradient-based techniques (e.g., Grad-CAM) were applied in image analysis.

Obraz zawierający tekst, zrzut ekranu, numer, diagram

Opis wygenerowany automatycznie

“**By far the most frequent ML task among the reviewed XAI papers is classification**, followed by regression and clustering. Among the used ML models, **deep neural networks are predominant, especially convolutional neural networks**. The second most used group of ML models are **tree-based models (decision and regression trees, random forest, and other types of tree ensembles)**. Interestingly, there is no substantial difference between the major ML models with regard to the ML task of their target application.”

Key challenges:

* Lack of consistent metrics for evaluating the effectiveness of explanations.
* Dependence on expert opinions and limitations in the number of studies with formal assessments of XAI quality.

1. <https://www.sciencedirect.com/science/article/pii/S0950705123000230>

**Explainable AI (XAI): A systematic meta-survey of Current challenges and Future opportunities**

**Scope of the study:** A meta-review of the literature focusing on the challenges and future research directions of XAI.

**About SHAP and LIME**

**SHAP (Shapley Additive Explanations)**:

* offers mathematically rigorous explanations by attributing feature importance in a way that maintains model consistency, ideal for high-stakes domains.
* popularity in interpretability
* **limitation in handling highly correlated features, making its utility in citation screening effective but computationally intensive on larger datasets**

**(Local Interpretable Model-agnostic Explanations): LIME**

* model-agnostic approach
* interpret complex models locally, which is advantageous in domain-specific citation screening where interpretations must align closely with user expertise.
* **Lack of stability and consistency in high-volume applications.**

**Key challenges:**

* Lack of consensus on the definitions of "explainability" and "interpretability."
* Limitations in the development of formal metrics for evaluating interpretability.
* High costs of developing and deploying XAI systems, especially in industry.
* Challenges in explaining AI model decisions in critical applications such as healthcare and security.

**Perspectives and motivations for XAI:**

* **Regulatory**: The need for transparency in light of regulations such as GDPR.
* **Scientific**: Understanding the patterns that ML models identify in data.
* **Industrial**: Building trust in AI models by improving their transparency.
* **Social**: Ensuring fairness and comprehensibility in model decisions.

1. Z slacka : <https://www.sciencedirect.com/science/article/abs/pii/S0169260722005429>

***"Explainable AI: A Systematic Literature Review Focusing on Healthcare"***

Analyzing 50 studies of XAI in healthcare.

Highlighting purpose of XAI in healthcare, with emphasis on importance of transparent AI models to foster trust and enable effective clinical integration. Besides, examining popular algorithms like SHAP & LIME.

Main challenges:

* Lack of clean interpretation about model’s decisions in a sample, which results in struggle with interpreting machine predictions.
* Lack of real-world data and limited involvement of medical experts in model development

Success:

* XAI is valuable in enhancing clinical decision-making by clarifying model predictions
* XAI in Medical image analysis, identifying critical features for accurate diagnosis:
  + SHAP in identifying image areas with critical features for diagnoses
  + LIME in local (single record) explanations