Linnéuniversitetet
Kalmar Växjö

Demystifying AI:

Five Techniques to Explain Model Decisions

Author 1: Alambuya Chelsea Peace
Author 2: Naomi Chapman
Supervisor: Amilcar Soares
Semester: HT24
Discipline: Software Technology
Course code: 2DV505

Abstract

As Artificial Intelligence (AI) systems become more integral in sensitive fields like healthcare, finance and law, the ability to understand their decision making process is crucial. This review explores five key studies in the field of Explainable AI (XAI) that tackle challenges in improving model interpretability for high-stakes fields. Key innovations include improved Shapley value approximations for feature interaction modelling, enhanced local explanations through G-LIME's use of global dataset patterns, and the selector-predictor-guider framework for coherent rationale extraction. Furthermore, Graph Convolutional Neural Networks (GCNNs) combined with Layer-wise Relevance Propagation (LRP) demonstrate stability in gene selection for cancer diagnosis, while studies on counterfactual explanations emphasize aligning machine-generated insights with human reasoning. Despite progress, challenges like computational demands, inconsistent evaluation metrics, and insufficient fairness focus remain. Future research should prioritize optimizing computational efficiency, developing consistent benchmarks, and addressing ethical considerations to build trust and reliability in AI systems.

Introduction

Artificial Intelligence (AI) models are being increasingly deployed in sensitive fields like healthcare, finance, judicial, and autonomous systems. AI models like deep neural networks are often described as "black boxes", as there is no real insight into their decision-making process. This makes the models hard to understand which reduces trust in their efficacy and impartiality, especially in the above mentioned critical fields. Explainable AI(XAI) is a field of research that attempts to illuminate the "black box" and some techniques have been developed to improve AI transparency.

This review examines recent advancements in XAI techniques, focusing on five papers that address major challenges in model interpretability and optimal feature selection. These papers explore managing feature dependencies, stabilizing explanations, and refining rationale generation. Together, these studies represent important steps toward making AI systems more interpretable and suitable for high-stakes applications, advancing transparency and reliability across diverse fields.

Literature Reviews

"Explaining individual predictions when features are dependent: More accurate approximations to Shapley values" by K. Aas, M. Jullum, A. Løland (2021)

Kernel SHAP is a tool used to explain predictions in complex models by assigning a "credit" to each feature using Shapley values from game theory. Shapley values measure each feature's contribution to a prediction by fairly distributing the "credit" according to all the potential feature combinations. For example, in a model predicting loan approvals, Kernel SHAP can explain how features like income, credit score, and employment history contribute to the final decision. However, it assumes features don't depend on each other, which can lead to inaccurate explanations when they are related, such as age and cholesterol levels in medical data.

To address this, the authors propose three new methods: a multivariate Gaussian model, a Gaussian copula for flexible modelling, and an empirical method that uses real data to capture feature relationships. They also suggest a combined approach that adapts based on the number of features. Tests on simulated and real data show these methods improve explanations by capturing feature interactions and perform better than the original Kernel SHAP when dependencies are present. The authors use Mean Absolute Error (MAE) to measure accuracy, showing that each new approach produces lower MAE values, indicating improved performance.

The study's strengths are its thorough testing and ability to produce fair explanations even with complex data. However, these methods require more computing power, which could limit their use on large datasets. Fields like finance and healthcare, where feature dependencies are common, would benefit most from these improvements.

"G-LIME: Statistical learning for local interpretations of deep neural networks using global priors" by X. Li, H. Xiong, X. Li, Z. Zhang, J. Liu, H. Jiang, Z. Chen and D. Dou (2023)

This study aims to improve the interpretability of Deep Neural Networks (DNNs) by enhancing Local Interpretable Model-agnostic Explanations (LIME), a popular method used to explain individual predictions in complex models. LIME creates simpler, interpretable models (like linear models) around each prediction, explaining the local behaviour of the complex model.

The authors present G-LIME, an approach that incorporates global dataset patterns to produce more stable and reliable explanations, addressing LIME's inconsistency due to random sampling. G-LIME builds on LIME using a Bayesian framework with ElasticNet regularization, combining L1 regularization to simplify explanations by focusing on fewer features and L2 regularization to capture broader dataset patterns. It also employs NormLIME (for consistency across predictions) and Least Angle Regression (LARS) to efficiently rank feature importance, making it adaptable to diverse data types.

Tests show that G-LIME provides more consistent and accurate explanations than LIME across various datasets, stabilizing explanations while remaining flexible for different data types. However, it requires more computational power and retains some randomness, limiting its scalability. Despite these challenges, G-LIME represents a significant improvement over LIME and, with refinement, could better handle adversarial inputs.

"Rationalizing predictions by adversarial information calibration" by L. Sha, O. Camburu and T. Lukasiewicz (2023)

The paper explores a model that combines a selector-predictor-guider framework with adversarial information calibration and a language model regularizer to enhance AI interpretability, particularly in safety-critical fields. The key objectives are to enhance rationale extraction by refining the selection of essential features through adversarial calibration, ensure coherence in natural language tasks via a language model regularizer, and maintain high predictive accuracy.

The framework is comprised of a selector to identify relevant features, a predictor to make predictions based solely on these features, and a guider to inform the selection process. The adversarial setup calibrates the information between these models, while the language model regularizer ensures coherent rationales. The results demonstrate significant improvements in rationale precision and recall across datasets for sentiment analysis, hate speech detection, and legal judgments, without compromising predictive accuracy.

The main strength of the study is it's relatively high interpretability and coherent rationale generation which is especially useful for applications in the legal and medical fields where transparency is required. However, the proposed model incurs a high computational cost and relies on the quality of the pre-trained language models.

"Stable feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation for biomarker discovery in breast cancer" by H. Chereda, A. Leha and T. Beißbarth (2024)

This paper explores enhancing stability and interpretability in gene selection for breast cancer prognosis. Graph Convolutional Neural Networks (GCNNs) are neural networks that work well with graph data, such as protein-protein interaction networks, which capture relationships between elements. Layer-wise Relevance Propagation (LRP) assigns "relevance" scores to features, helping to explain predictions by highlighting the most relevant genes. For example, in cancer diagnosis, LRP can identify genes contributing most to the prediction. The authors combine GCNNs with LRP to improve feature stability and biological relevance, comparing their approach to GCNN+SHAP (using Shapley values), Multi-Layer Perceptron (MLP), and Random Forest (RF) models.

Results show that GCNN+LRP produces the most stable and interpretable gene sets, outperforming others in pathway relevance. While GCNN+SHAP offers high classification impact, its stability falls short. The GCNN+LRP model's higher computational demand stems from the complexity of processing graph data and computing relevance scores, which could limit scalability. Strengths of this study include thorough comparisons and its relevance to biomarker discovery. Future research could focus on optimizing this approach to balance stability, interpretability, and efficiency for larger datasets.

"Counterfactual explanations for misclassified images: How human and machine explanations differ" by E. Delaney, A. Pakrashi, D. Greene and M. T. Keane (2023)

This paper introduces Counterfactual explanations which provide what-if explanations to show how the model prediction would change if the input features had been different in some way. The paper aims to investigate how human-generated counterfactual explanations compare to those produced by various computational methods in image classification tasks.

The study uses two image datasets to collect explanations from humans and benchmark counterfactual algorithms like Min-Edit, CEM-PN, CEGP, and Revise. The results reveal that humans tend to make larger, semantically meaningful edits that align with prototypes of the counterfactual class, while computational methods typically use minimal edits that may not capture the representativeness or prototypicality humans utilize.

Strengths of this study include its user-centered approach, highlighting the differences in explanation goals between humans and machines, which can guide future algorithm improvements. However, a limitation is the small sample size in one of the dataset pilot, which might restrict the generalizability of findings to more complex image datasets

Discussion

These studies offer valuable insights into XAI, especially in critical fields like healthcare, law, and finance, where reliable explanations are essential. A key focus across the papers is creating stable, dependable explanations by managing feature dependencies and generating context-based rationales. This trend in XAI reflects a move toward models that are accurate and trustworthy. In sensitive areas, unreliable AI predictions could lead to serious issues, such as misdiagnoses or poor financial advice. These XAI methods aim to make AI safer by making predictions clearer and easier to trust.

A common thread between all papers is the significance of input features, how they are selected, how they interact with each other and how they are manipulated to improve the explanations. The paper exploring more accurate approximations to Shapely values highlight that accounting for feature dependencies provide more accurate explanations as compared to when features are assumed to be independent of each other [1]. The study implementing the G-LIME approach focuses on using fewer features selected by efficiently ranking their importance [2]. In the same vein, with the selector-predictor-guide framework, the selector model identifies the relevant features informed by the guider model [3] while with LRP features are assigned a relevance score to identify the most important ones [4]. And lastly Counterfactual explanations highlight which features contribute most to a model's prediction [5].

The central goal among the papers is achieving stability and reliability in explanations. Both G-LIME and the adversarial calibration framework work to provide consistent explanations across different data conditions [2] [3]. G-LIME achieves this by drawing on patterns within the dataset, while the adversarial calibration framework uses the three-part model to refine explanations, particularly for text-based tasks. Although they take different approaches, both methods seek to avoid unpredictable variations in explanations, which helps build user trust.

The focus on stability and reliability in these studies shows how XAI is moving from being a theoretical idea to something that can be practically used. This shift is essential for using AI in critical areas, where decisions can have serious real life consequences. For example, GCNN+LRP helps make gene selection for cancer diagnosis more reliable, improving the consistency of results in medical tools [4]. Similarly, G-LIME provides stable and clear explanations, which are vital for financial analysts who need dependable insights [2]. These methods not only improve the accuracy of AI systems but also make them more trustworthy for users, encouraging their use in real-world settings.

Another important connection between these studies is their focus on bridging the gap between how human and machine explanations. Delaney et al.'s work on counterfactual explanations highlights the need to create AI outputs that match how people think, making them easier to understand and use [5]. This is especially important in areas like law, where professionals need clear reasons behind AI decisions. Together, these papers help move XAI toward systems that are not just technically effective but also practical and easy for people to trust and use in everyday life.

The reviewed studies highlight important advancements in Explainable AI (XAI), but they also reveal some key gaps. One major issue is the high computational cost of methods which makes them hard to use with large datasets or in real-time applications. Optimizing these methods for efficiency while keeping their accuracy is an area for improvement. Another challenge is the lack of standard evaluation metrics across the studies. Each paper uses different measures like stability, accuracy, or recall, making it difficult to compare methods directly. Just within the investigation of counterfactual explanations it is highlighted that explanation goals are not consistent in the research [5]. Developing consistent benchmarks could help address these inconsistency issues.

Another gap is the limited focus on how users interact with these explanations. While Delaney et al.'s work considers human perspectives [5], the other four papers don't explore how users interpret or trust the explanations. Additionally, many methods are designed for specific fields, like healthcare or text analysis, with little evidence of how they work in other domains. Finally, fairness and bias in AI are not deeply explored in these studies, even though they are essential for ensuring equitable and responsible use of AI systems. Future research should aim to develop methods that are efficient, adaptable across domains, and designed to promote fairness and transparency.

Conclusion

The studies showcase significant progress in the XAI field, honing in on creating stable, reliable and context-aware explanations which is becoming increasingly important as AI is adopted in a broad spectrum of domains. The techniques brought forward aim to increase the trustworthiness of AI systems in different ways that begin with prime feature selection and ends with clear and consistent explanations that align with human reasoning.

However, challenges still exist. The main one being the high computational cost of analysing larger datasets which limits the scope of how different techniques can be tested and applied. Additionally, there is limited user interaction and this is particularly problematic as techniques are being designed to be applied in different fields and yet the domain experts in the proposed fields are not a part of the experimentation. Also missing from the papers is a significant examination of fairness and bias which is critical if AI systems are to be implemented in various domains that will impact people's lives.

Further research should be directed towards reducing the computational cost of processing larger datasets, establishing standardized evaluation metrics and goals and incorporate fairness while working with AI systems. XAI as field is critical in bridging the gap between human and machine understanding which will solidify trust and reliability in AI systems

References

[1] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," *Artificial Intelligence*, vol. 298, p. 103502, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000539

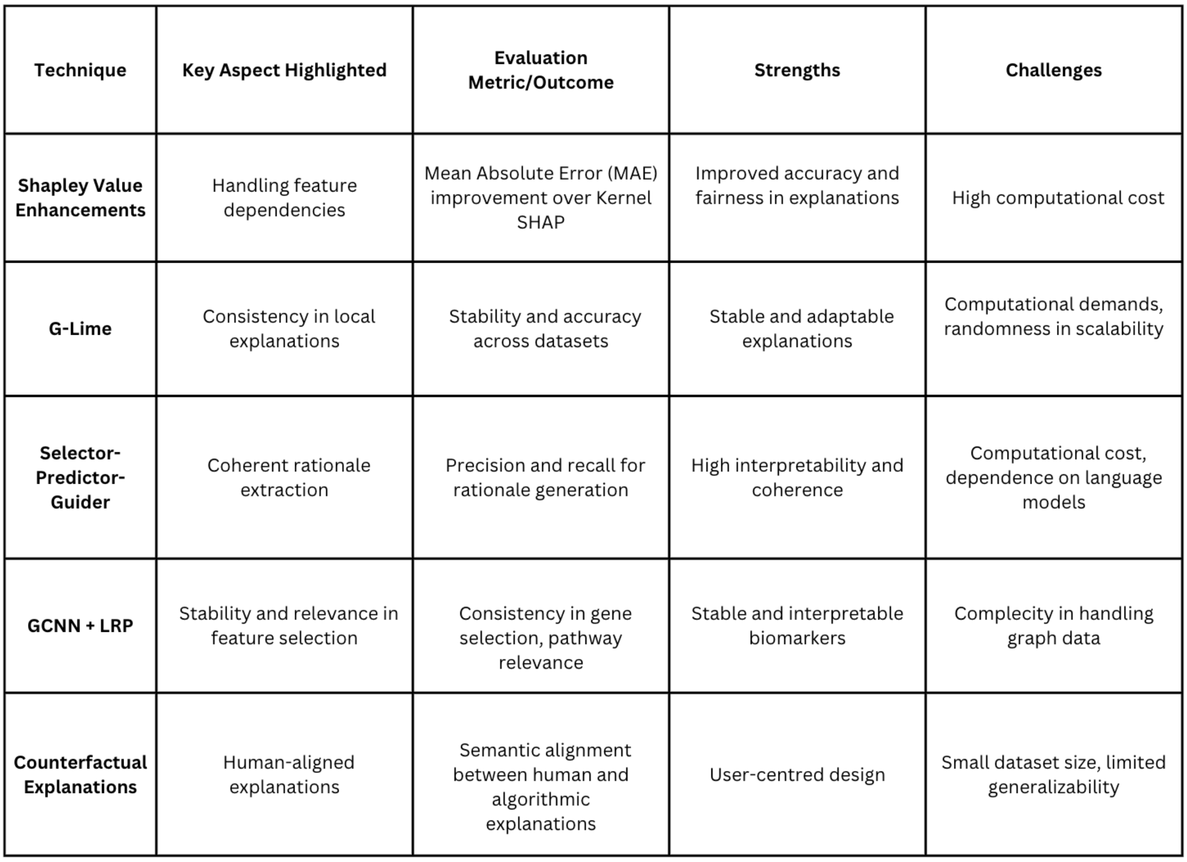
[2] X. Li, H. Xiong, X. Li, X. Zhang, J. Liu, H. Jiang, Z. Chen, and D. Dou, "G-lime: Statistical learning for local interpretations of deep neural networks using global priors," *Artificial Intelligence*, vol. 314, p. 103823, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370222001631

[3] L. Sha, O.-M. Camburu, and T. Lukasiewicz, "Rationalizing predictions by adversarial information calibration," *Artificial Intelligence*, vol. 315, p. 103828, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370222001680

[4] H. Chereda, A. Leha, and T. Beißbarth, "Stable feature selection utilizing graph convolutional neural network and layer-wise relevance propagation for biomarker discovery in breast cancer," *Artificial Intelligence in Medicine*, vol. 151, p. 102840, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365724000824

[5] E. Delaney, A. Pakrashi, D. Greene, and M. T. Keane, "Counterfactual explanations for misclassified images: How human and machine explanations differ," *Artificial Intelligence*, vol. 324, p. 103995, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370223001418

Appendix



Technique	Key Aspect Highlighted	Evaluation Metric/Outcome	Strengths	Challenges
Shapley Value Enhancements	Handling feature dependencies	Mean Absolute Error (MAE) improvement over Kernel SHAP	Improved accuracy and fairness in explanations	High computational cost
G-Lime	Consistency in local explanations	Stability and accuracy across datasets	Stable and adaptable explanations	Computational demands, randomness in scalability
Selector-Predictor-Guider	Coherent rationale extraction	Precision and recall for rationale generation	High interpretability and coherence	Computational cost, dependence on language models
GCNN + LRP	Stability and relevance in feature selection	Consistency in gene selection, pathway relevance	Stable and interpretable biomarkers	Complecity in handling graph data
Counterfactual Explanations	Human-aligned explanations	Semantic alignment between human and algorithmic explanations	User-centred design	Small dataset size, limited generalizability

Figure 1: Table: Summary of Techniques, Key Aspects, Evaluation Metrics, Strengths, and Challenges