Understanding explainable artificial intelligence techniques: a comparative analysis for practical application

Shweta Bhatnagar, Rashmi Agrawal
School of Computer Applications, Manav Rachna International Institute of Research Studies (MRIIRS), Faridabad, India

Article Info

Article history:

Received Feb 22, 2024
Revised Apr 10, 2024
Accepted May 17, 2024

Keywords:

Explainable artificial intelligence
Explainable artificial intelligence models
Explainable artificial intelligence techniques
Local interpretable model-agnostic explanations
Shapley additive explanations

ABSTRACT

Explainable artificial intelligence (XAI) uses artificial intelligence (AI) tools and techniques to build interpretability in black-box algorithms. XAI methods are classified based on their purpose (pre-model, in-model, and post-model), scope (local or global), and usability (model-agnostic and model-specific). XAI methods and techniques were summarized in this paper with real-life examples of XAI applications. Local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP) methods were applied to the moral dataset to compare the performance outcomes of these two methods. Through this study, it was found that XAI algorithms can be custom-built for enhanced model-specific explanations. There are several limitations to using only one method of XAI and a combination of techniques gives complete insight for all stakeholders.

Corresponding Author:

Shweta Bhatnagar
School of Computer Applications, Manav Rachna International Institute of Research Studies (MRIIRS)
Faridabad, Haryana, India
Email: shwetabhatnagar99@gmail.com

1. INTRODUCTION

Explainable artificial intelligence (XAI) is a branch of artificial intelligence (AI) that develops AI systems that provide clear and transparent explanations for their decisions and predictions. XAI converts black-box AI systems into human understandable interpretations. In fields, such as finance, healthcare, and the criminal justice system, the decisions made by AI systems have serious consequences, requiring transparency through audit. XAI addresses this concern by developing AI systems that provide explanations for their decisions that are clear, concise, and easily understood by humans. XAI makes AI more trustworthy, transparent, and accountable. XAI can be seen as a bridge between the mathematical and statistical foundations of AI models and the human-understandable explanations required by the people who use or are affected by those models.

XAI explanations are classified as model-agnostic methods that provide explanations for the decisions made by any AI model, and model-specific methods that are tailored to the inner workings of specific AI models. Model-agnostic methods provide explanations for the predictions made by any AI model, regardless of its architecture or internal workings. Examples of model-agnostic methods include local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP). These methods can be applied to any model. Model-specific methods, on the other hand, are tailored to the inner workings of specific AI models and provide explanations that are specific to those models. Examples of model-specific methods include layer-wise relevance propagation (LRP) and gradient-based methods. These methods explain the contribution of each feature to the final prediction. They help explain how the model decides and the decision-making process.

Another approach to XAI classification is based on the performance of the model on a specific dataset. These are called local explanations and global explanations. Local explanations help in describing the reasons for a decision in case of a specific instance [1], while the global explanation helps in clarifying the feature importance of a complete model [2]. SHAP gives global explanations while LIME is only focused on local explanations.

XAI explanations are also classified as pre-model, in-model, and post-model techniques, where the use of techniques is based on the stage of model development for decision-making. Pre-model is closely related to the data interpretability and analysis. In-model techniques explain the intrinsic workings of the model. Post-model techniques share post-hoc insights to find ways of improving the decision through yet unknown findings [3].

In this paper, an exploratory study is conducted to understand XAI techniques, and tools applicable to real-life problems. The paper is divided into four parts. Section 1, discusses the introduction, followed by section 2 summarizing popular XAI techniques and tools. Section 3 discusses the findings on SHAP and LIME for moral datasets. Section 4 concludes with future work.

2. XAI TECHNIQUES AND TOOLS

The goal of XAI is to make AI models more trustworthy, transparent, and accountable [4]. Appropriate XAI tools can be built using a combination of techniques such as feature importance, local explanations, sensitivity analysis, and visualizations for the interpretability of AI models. XAI techniques used for improving explanations are explained in Table 1. Apart from these techniques, regular deep-learning algorithms can also be used for building transparent models. Popular XAI model-agnostic and model-specific methods used for explainability are compared in Table 2. Examples of XAI real-life applications published last year and the tools and techniques used in each application are shared in Table 3.

Table 1. XAI Techniques for explainability

Xai techniques	Technique description	Method	Scope of explanation
Feature importance [5]	Change in output for change in the feature value input	Model-agnostic	Local and global
Visualizations	Plots and graphs	Model-specific	Local and global
Sensitivity analysis [6]	Perturbation of data to measure impact	Model-agnostic	Global
Counterfactual explanations [7]	Alternate explanations	Model-specific	Global

Table 2. XAI tools for model-agnostic and model-specific explanations

XAI methods	Method description	Method	Scope of explanation
LIME [8]	Predictions on individual instances	Model-agnostic	Local
SHAP [9]	Feature importance of model	Model-agnostic	Local and global
Captum [10]	A library in PyTorch	Model-specific	Customizable
TensorFlow [11]	An open-source library	Model-specific	Customizable

Table 3. Real-life applications of XAI

Real-life applications of XAI	XAI tools and techniques used
Cancer detection [12]	Visualizations (partial dependence plot), feature importance, and SHAP
Visitor arrival forecast and reputation assessment [13]	Feature importance (PCA) and LIME
Root cause detection of disease [14]	Counterfactual description
Alzheimer’s disease [15]	SHAP, LIME, and deep learning algorithms
Medical text processing [16]	Gradient-weighted class activation mapping (grad-CAM) visualizations
Energy management [17]	Silhouette coefficients and PCA technique
Purchase prediction [18]	SHAP
Stress classification [19]	SHAP
Osteoporosis risk prediction [20]	SHAP, LIME, QLattice, and feature importance
EEG-based activity recognition [21]	LIME
Seizure-detection [22]	SHAP

The choice of a specific tool or technique best suited to a problem requires choosing XAI that best meets the requirements of stakeholders for its explanation. It is a combination of solutions that can address biases and fairness issues holistically for better and ethical decision-making. There are some limitations of existing XAI tools described:

- Increased computational time: building explainability can include computational time, which can affect the performance measures for algorithms.

- Lack of standards: missing standards for explainability requirements even by model-agnostic models causes ambiguity in the performance of such techniques [23].
- Limited explanation: model-agnostic XAI methods like LIME and SHAP provide isolated explanations. They behave differently for different datasets [24].
- Exclusive of actual decision: XAI methods do not form part of the decision-making process, but only provide explanations. Therefore, they are exclusive of any impact and do not guarantee fair algorithms.

3. XAI FOR MORAL DATASETS

3.1. Dataset description

The Kaggle dataset named ‘sex-differences-in-moral-judgements-67-countries with 11,969 responses from 67 countries was used in the analysis. The dataset is based on moral foundational theory (MFT) with 5 moral foundations of psychology namely, care vs harm, fairness vs cheating, loyalty/ingroup vs betrayal, authority vs subversion, and purity vs degradation [25]. MFT is mapped with the life satisfaction index (LSI) of countries to understand each factor contribution. LSI is calculated based on surveys within countries and bases its values on several factors such as mood, zest for life, aging, and other parameters. The gradient-boosting regressor model was used in both SHAP and LIME to explain complex relationships between MFT and LSI.

3.2. Explainability using SHAP

SHAP uses game theory to identify important features in AI models. The results are local for instance predictions and global for the whole model. It is model-agnostic and can be applied to any model. SHAP was implemented on the MFT dataset to find pre-model explainability concerning country-wise LSI. The summary plot visualizing the ranks of features through SHAP is shown in Figure 1. It shows that the fairness vs cheating and purity vs degradation values are more relevant for explaining the model globally. Although it does not share any insights on the kind of relationships, such as linear or non-linear. The gradient boosting regressor model used in SHAP shows that the error rate is high and the explained variance is low, as seen in Table 4.

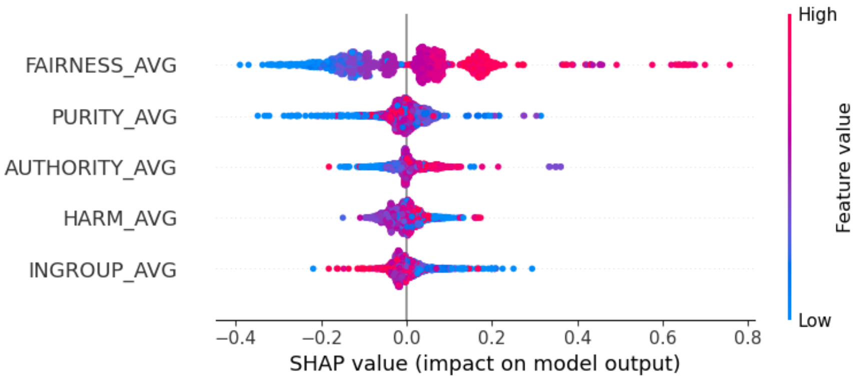


Figure 1. Results of SHAP on MFT database

Table 4. Comparison of performance of SHAP and LIME

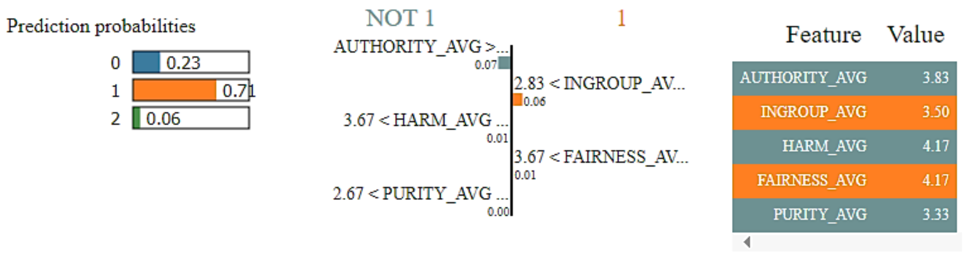
Tool	MSE	MAE	R-squared	Explained variance
SHAP	37.42619104472565	6.115310307760435	0.08657594621162412	0.08658230785735055
LIME	0.02608564859807469	0.1208246671695897	0.02090236330133688	0.03257132885542591

3.3. Explainability using LIME

LIME is an open-source local model-agnostic tool for explanations of black-box algorithms. LIME makes explanations on data levels. It explains the predicted instance by minimizing the sum of the loss function and interpretability complexity, given in (1). An instance-specific explanation using linear regression for pre-model MFT explanation of LSI using LIME is shown in Figure 2. The gradient boosting regressor model used in LIME showed that the error rates are low for LIME, but the explanations are restricted, as seen in Table 4.

$$E(x) = L(f, g, \pi_x) + \Omega(g) \quad (1)$$

Where $L(f, g, \pi_x)$ is the loss or fidelity function, Ω(g) is the complexity of interpretation, f is the model around x, g represents a surrogate machine-learning model, x is the instance, and π_x represents the perturbed instances.



Feature	Value
AUTHORITY_AVG	3.83
INGROUP_AVG	3.50
HARM_AVG	4.17
FAIRNESS_AVG	4.17
PURITY_AVG	3.33

Figure 2. Results of LIME on MFT database

3.4. Comparative analysis of SHAP and LIME

A comparison of SHAP and LIME performance is given in Table 4 and it validates that:

- SHAP algorithm gives a global explanation to models while LIME gives local explanations
- LIME has higher accuracy for each prediction but limited insight into model
- Both SHAP and LIME are model-agnostic and can be applied to any model
- Both LIME and SHAP leave scope for post-hoc model explanations to answer "*what more can be found*" [26]
- For greater insights on models, a combination including other techniques is desirable [27]

4. CONCLUSION

Explaining the AI models is very important from an ethical perspective of using AI. Explainable models that give an insight into the data and features are developed using XAI techniques and methods. Finding the right tool and technique depends on the size and type of data and model. Model-agnostic methods like SHAP or LIME are popularly used on a wide variety of data and models in real-life applications of XAI. Research revealed that a combination of techniques and methods used for explaining models for a diverse group of stakeholders instead of only one technique had greater results. XAI methods have limitations as independent techniques and therefore should be customized as a combination of methods for each problem at hand.

As part of future work, model-specific models can be developed using techniques and methods on any suitable database. XAI can be used for designing models as part of the pre-model stage. A combination of techniques to develop decision rules that are transparent for businesses can be rewarding.

REFERENCES

[1] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: opportunities and challenges of XAI-based model improvement," *Information Fusion*, vol. 92, pp. 154–176, 2023, doi: 10.1016/j.inffus.2022.11.013.

[2] V. Keppeler, M. Lederer, and U. A. Leucht, "Explainable artificial intelligence," *Encyclopedia of Data Science and Machine Learning*, pp. 1667–1684, 2022, doi: 10.4018/978-1-7998-9220-5.ch100.

[3] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: a survey on methods and metrics," *Electronics (Switzerland)*, vol. 8, no. 8, 2019, doi: 10.3390/electronics8080832.

[4] M. Langer *et al.*, "What do we want from explainable artificial intelligence (XAI)? a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artificial Intelligence*, vol. 296, 2021, doi: 10.1016/j.artint.2021.103473.

[5] K. Främling, "Feature importance versus feature influence and what it signifies for explainable AI," *Communications in Computer and Information Science*, vol. 1901, pp. 241–259, 2023, doi: 10.1007/978-3-031-44064-9_14.

[6] D. Alvarez-Melis and T. S. Jaakkola, "A causal framework for explaining the predictions of black-box sequence-to-sequence models," *Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2017, pp. 412–421, doi: 10.18653/v1/d17-1042.

[7] J. Pearl, *Causality*, Cambridge University Press, 2009.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' explaining the predictions of any classifier," *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.

[9] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *A NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4766–4775, doi: 10.5555/3295222.3295230.

[10] N. Kokhlikyan *et al.*, "Captum: a unified and generic model interpretability library for PyTorch," *arXiv:2009.07896*, 2020, doi: 10.48550/arXiv.2009.07896.

[11] I. Hull, "TensorFlow 2," em *Machine Learning for Economics and Finance in TensorFlow 2*, Springer Link, 2020, pp. 1-59.

[12] T. Khater, S. Ansari, S. Mahmoud, A. Hussain, and H. Tawfik, "Skin cancer classification using explainable artificial intelligence on pre-extracted image features," *Intelligent Systems with Applications*, vol. 20, 2023, doi: 10.1016/j.iswa.2023.200275.

[13] E. Collini, P. Nesi, and G. Pantaleo, "Reputation assessment and visitor arrival forecasts for data driven tourism attractions assessment," *Online Social Networks and Media*, vol. 37–38, 2023, doi: 10.1016/j.osnem.2023.100274.

[14] E. V. Strobl, "Counterfactual formulation of patient-specific root causes of disease," *Journal of Biomedical Informatics*, vol. 150, 2024, doi: 10.1016/j.jbi.2024.104585.

[15] G. Loveleen, B. Mohan, B. S. Shikhar, J. Nz, M. Shorfuzzaman, and M. Masud, "Explanation-driven HCI model to examine the mini-mental state for Alzheimer's disease," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 2, 2023, doi: 10.1145/3527174.

[16] H. Zhang and K. Ogasawara, "Grad-CAM-based explainable artificial intelligence related to medical text processing," *Bioengineering*, vol. 10, no. 9, 2023, doi: 10.3390/bioengineering10091070.

[17] D. P. Panagoulias, E. Sarmas, V. Marinakis, M. Virvou, G. A. Tsihrintzis, and H. Doukas, "Intelligent decision support for energy management: a methodology for tailored explainability of artificial intelligence analytics," *Electronics (Switzerland)*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214430.

[18] B. Predić, M. Ćirić, and L. Stoimenov, "Business purchase prediction based on XAI and LSTM neural networks," *Electronics (Switzerland)*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214510.

[19] J. Tervonen, J. Närväinen, J. Mäntyjärvi, and K. Pettersson, "Explainable stress type classification captures physiologically relevant responses in the Maastricht acute stress test," *Frontiers in Neuroergonomics*, vol. 4, 2023, doi: 10.3389/fnrgo.2023.1294286.

[20] V. V. Khanna *et al.*, "A decision support system for osteoporosis risk prediction using machine learning and explainable artificial intelligence," *Heliyon*, vol. 9, no. 12, 2023, doi: 10.1016/j.heliyon.2023.e22456.

[21] I. Hussain *et al.*, "An Explainable EEG-based human activity recognition model using machine-learning approach and LIME," *Sensors*, vol. 23, no. 17, 2023, doi: 10.3390/s23177452.

[22] J. C. Vieira, L. A. Guedes, M. R. Santos, and I. Sanchez-Gendriz, "Using explainable artificial intelligence to obtain efficient seizure-detection models based on electroencephalography signals," *Sensors*, vol. 23, no. 24, 2023, doi: 10.3390/s23249871.

[23] H. L. B. Chia, "The emergence and need for explainable AI," *Advances in Engineering Innovation*, vol. 3, no. 1, pp. 1–4, 2023, doi: 10.54254/2977-3903/3/2023023.

[24] C. Steging, S. Renooij, and B. Verheij, "Rationale discovery and explainable AI," *Frontiers in Artificial Intelligence and Applications*, vol. 346, pp. 225–234, 2021, doi: 10.3233/FAIA210341.

[25] M. Atari, M. H. C. Lai, and M. Dehghani, "Sex differences in moral judgements across 67 countries," *Proceedings: Biological Sciences*, 2020, doi: 10.1098/rspb.2020.1201

[26] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186, doi: 10.1145/3375627.3375830.

[27] M. Nagahisarchoghaei, M. M. Karimi, S. Rahimi, L. Cummins, and G. Ghanbari, "Generative local interpretable model-agnostic explanations," *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*, vol. 36, 2023, doi: 10.32473/flairs.36.133378.

BIOGRAPHIES OF AUTHORS

Shweta Bhatnagar is an academician with an experience over a decade in education administration at various public and private organizations. She was a JRF at ICPR and is a Ph.D. student at Manav Rachna International Institute of Research Studies (MRIIRS). Her field of research is education technology. She can be contacted at email: shwetabhatnagar99@gmail.com.

Dr. Rashmi Agrawal is Ph.D. and UGC-NET qualified with 20 years of experience in teaching and research, working as Professor in Department of Computer Applications, Manav Rachna International Institute of Research Studies, Faridabad, India. She is associated with various professional bodies in different capacities, life member of Computer Society of India and senior member of IEEE, she is book series editor of Innovations in Big Data and Machine Learning, CRC Press, Taylor and Francis group, USA and Advances in Cybersecurity in Wiley. She has authored/co-authored many research papers in peer reviewed national/international journals and conferences which are SCI/SCIE/ESCI/SCOPUS indexed. She has also edited/authored books with national/international publishers (Springer, Elsevier, IGI Global, Apple Academic Press, and CRC Press) and contributed chapters in books. Currently she is guiding Ph.D. scholars in sentiment analysis, educational data mining, internet of things, brain computer interface and natural language processing. She is Associate Editor in Journal of Engineering and Applied Sciences and Array Journal, Elsevier. She can be contacted at email: Drrashmiagrawal78@gmail.com.