Research Reports on Computer Science
https://ojs.wiserpub.com/index.php/RRCS

Article

A Comparative Analysis of Model Agnostic Techniques for Explainable Artificial Intelligence

Yifei Wang

School of Information, University of California, Berkeley, Berkeley, CA, USA
E-mail: sarahwang688@berkeley.edu

Received: 13 April 2024; **Revised:** 1 July 2024; **Accepted:** 22 July 2024

Abstract: Explainable Artificial Intelligence (XAI) has become essential as AI systems increasingly influence critical domains, demanding transparency for trust and validation. This paper presents a comparative analysis of prominent model agnostic techniques designed to provide interpretability irrespective of the underlying model architecture. We explore Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) plots, and Anchors. Our analysis focuses on several criteria including interpretative clarity, computational efficiency, scalability, and user-friendliness. Results indicate significant differences in the applicability of each technique depending on the complexity and type of data, highlighting SHAP and LIME for their robustness and detailed output, whereas PDP and ICE are noted for their simplicity in usage and interpretation. The study emphasizes the importance of context in choosing appropriate XAI techniques and suggests directions for future research to enhance the efficacy of model agnostic approaches in explainability. This work contributes to a deeper understanding of how different XAI techniques can be effectively deployed in practice, guiding developers and researchers in making informed decisions about implementing AI transparency.

Keywords: Artificial Intelligence, Explainable Artificial Intelligence, machine learning, AI techniques

1. Introduction

The increasing integration of Artificial Intelligence (AI) systems in critical sectors such as healthcare, finance, and autonomous driving demands not only high accuracy but also a clear understanding of how decisions are made. Explainable AI (XAI) seeks to bridge the gap between AI performance and human comprehension, ensuring that AI decisions are transparent, trustworthy, and easy to interpret. This need is underscored by regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR), which includes provisions for the right to explanation of automated decisions [1].

AI models are often considered "black boxes" because their decision-making processes are not easily understood by humans. This lack of transparency can be problematic, especially in high-stakes domains where understanding the rationale behind a decision is crucial. Model agnostic techniques are particularly valuable in this context because they provide insights into AI models without the need to access or alter the underlying algorithms. This universality makes them suitable for a wide range of industries and applications. For example, in healthcare, understanding the decision-making process of AI can help clinicians validate diagnoses suggested by AI [2]. In finance, where AI is used to assess creditworthiness or

DOI: https://doi.org/10.37256/rrcs.3220244750

manage investments, explainability can help identify biases and ensure fair lending practices [3]. Autonomous driving also benefits from explainable models, as manufacturers need to demonstrate how vehicles make decisions in critical situations to regulators and the public [4].

This paper focuses on a comparative analysis of prominent model agnostic techniques: Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) plots, and Anchors. These techniques have been chosen for their widespread use and potential applicability across various AI applications. By evaluating these techniques across different datasets and AI models, this study aims to identify the strengths and limitations of each, providing a guideline on their practical deployment in industry-specific applications.

The objective of this analysis is threefold: to enhance the understanding of each technique's operational mechanism, to evaluate their performance in real-world scenarios, and to aid stakeholders in selecting the most appropriate XAI technique based on specific needs. The contributions of this paper are intended to help pave the way for more transparent, understandable, and thus more ethically aligned AI systems across industries.

2. Background

Explainable AI (XAI) has emerged as a critical field in response to the growing complexity and ubiquity of AI models across various industries. The need for XAI stems from both ethical imperatives and practical necessities, as decision-making processes become increasingly automated and influential in sensitive domains. The foundational concept behind XAI is not merely to make AI systems transparent but to make their operations understandable and justifiable to human users [5].

2.1 *XAI Techniques: Model-Specific vs. Model-Agnostic*

XAI techniques are typically categorized into two types: model-specific and model-agnostic. Model-specific techniques are designed to work with particular types of models, exploiting their internal mechanisms to explain their behavior. For instance, attention mechanisms in neural networks provide insights into which parts of input data are being prioritized [6]. However, these techniques often lack flexibility as they are not applicable to models with different architectures.

Model-agnostic techniques, by contrast, are designed to be used with any model, providing flexibility and wide applicability. This universality makes them particularly valuable in industries where various AI models are employed. For example, in healthcare, model agnostic tools can help explain AI-driven predictive models for patient outcomes, regardless of whether they are based on logistic regression or complex neural networks [2]. Similarly, in finance, where models range from simple linear models for credit scoring to deep learning models for algorithmic trading, model-agnostic explanations aid in ensuring compliance and transparency across differing methodologies [7].

2.2 *Importance of Model Agnostic Techniques in Industry*

The appeal of model agnostic techniques in industry is largely due to their adaptability and ease of integration. In the automotive industry, for instance, where AI models predict vehicle failures or optimize logistics, the ability to apply a single explanatory technique across different AI models streamlines the process of validation and regulatory compliance [4]. In the field of retail, companies use a variety of AI models to forecast sales, personalize marketing, and manage supply chains. Here, model agnostic explanations assist in making these AI-driven decisions understandable to managers and stakeholders who may not have deep technical knowledge [3].

Furthermore, the implementation of model-agnostic XAI techniques aligns with legal frameworks such as the GDPR, which requires explanations of decisions made by AI systems affecting EU citizens. This legal requirement makes it essential for businesses operating in or with the EU to adopt XAI practices that can be applied regardless of the underlying AI technology [1].

2.3 *Key Model Agnostic Techniques*

Several model-agnostic techniques have gained prominence, each with its methodology and application domain: LIME (Local Interpretable Model-agnostic Explanations) provides local explanations for any classifier's predictions, making it suitable for use in medical diagnosis systems [7]. SHAP (SHapley Additive exPlanations), which utilizes game theory to explain the output of any machine learning model, is frequently used in finance to decompose the allocation of credit or risk [8]. Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots graphically depict the effects of the input variables on the predicted outcome, widely applicable from marketing to supply chain analytics. Anchors explain individual predictions based on the conditions that sufficiently anchor the prediction locally, useful in scenarios requiring highly reliable explanations, such as autonomous vehicle navigation.

3. Model agnostic techniques analysed

The effectiveness of explainable AI (XAI) hinges on the robustness and flexibility of its underlying techniques. Model agnostic approaches, notable for their versatility across different types of AI models, provide a broad toolkit for interpretability. Here, we delve into five prominent techniques, exploring their mechanisms, advantages, and practical applications in more detail.

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

LIME works by approximating the local decision boundary of the model around a prediction. It perturbs the input data, generating new data points, and observes how the predictions change. LIME then fits a simple, interpretable model (like a linear regression) to these new samples, weighting them according to their proximity to the original data point, to explain the prediction locally [7].

In healthcare, LIME helps demystify complex models used for predicting patient outcomes by highlighting influential factors in individual predictions. This can aid doctors in understanding the model's reasoning, potentially improving patient trust and adherence to treatment plans.

The process starts with the data point for which an explanation is needed. The model generates perturbations (slightly modified versions) of the original data point to create a dataset of neighbors around the original point. It then uses the black-box model to predict outcomes for each of the perturbed data points. Weights are assigned to each perturbed point based on its distance from the original data point, with closer points getting higher weights. Researchers typically then use a simple, interpretable model (e.g., linear regression) to fit the weighted perturbed data points. This model approximates the decision boundary of the black-box model locally around the original data point. The coefficients of the interpretable model provide an explanation of how the features of the original data point contribute to its prediction.

This architecture in Figure 1 below highlights the steps LIME uses to generate a local explanation, making the behavior of complex machine learning models more understandable in specific instances.

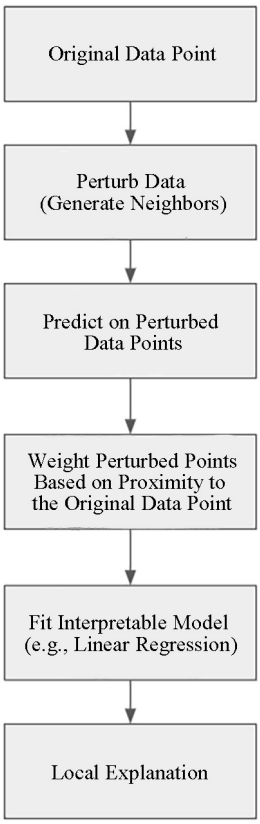

Figure 1. Architecture diagram of LIME

3.2 *SHapley Additive exPlanations (SHAP)*

SHAP values are derived from game theory, specifically the Shapley values, which are a method to fairly distribute the payout (prediction) among the players (features). SHAP values measure the contribution of each feature to the prediction by computing the change in the prediction when a feature is added to a subset of features, averaged over all possible subsets [8].

In finance, SHAP is used to explain individual loan approvals or rejections by quantifying the contribution of factors like income, credit history, and debt. This transparency is crucial for regulatory compliance and for improving the fairness of automated financial decisions.

The SHAP (SHapley Additive exPlanations) model architecture in Figure 2 begins with the **Input Data**, which consists of the dataset containing various features. This data is fed into the **Model**, a machine learning algorithm that generates **Predictions**. To interpret these predictions, the **SHAP Explainer** is used. The explainer employs Shapley values to assess the contribution of each feature to the model's output. The calculation process involves considering all possible subsets of features to determine each feature's marginal contribution, resulting in **Shapley Value Calculations** for each feature. These calculations yield **SHAP Values**, which quantify the importance of individual features in the model's predictions. The results are then utilized for **Local Explanations**, providing insights into individual predictions, **Global Explanations**, highlighting overall feature importance, and **Visualizations** such as graphs and plots, facilitating a comprehensive understanding of the model's behavior and enhancing transparency, trust, and compliance in AI applications.

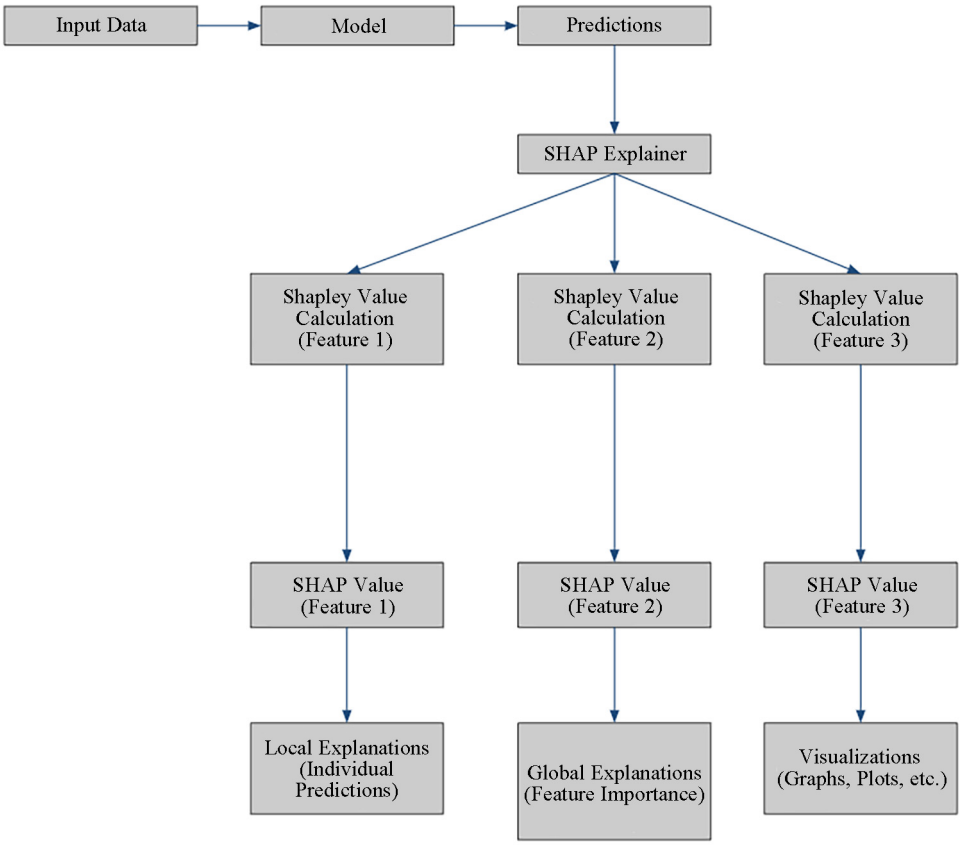

Figure 2. Architecture diagram of SHAP

3.3 *Partial Dependence Plots (PDP)*

PDPs show the marginal effect of one or two features on the predicted outcome. This is achieved by varying the feature(s) of interest across their range while keeping all other features constant at their average values, thus illustrating how the target prediction changes with changes in the input features [9].

Retail companies use PDPs to understand how price adjustments or changes in product features might affect sales volumes, assisting in strategic decision-making and pricing optimization.

3.4 *Individual Conditional Expectation (ICE) Plots*

ICE plots enhance the information provided by PDPs by plotting the dependency of the prediction on a feature for individual instances. Unlike PDPs that offer an average effect, ICE plots provide separate lines for each instance to show how the prediction changes with varying feature values, highlighting heterogeneity across the dataset [10].

In real estate, ICE plots can demonstrate how varying levels of renovations affect house prices on an individual basis, which can be instrumental for real estate agents and investors when making property enhancements.

3.5 *Anchors*

Anchors provide explanations by identifying "if-then" rules. These rules explain the conditions under which the same prediction would always be made, regardless of changes in other features. Anchors are particularly useful for explaining predictions in cases where certain features strongly anchor the output, providing a high level of certainty [11].

In autonomous driving, Anchors can clarify under what specific conditions (like weather, speed, and road type) the vehicle would decide to take a particular action, such as braking or changing lanes, enhancing the safety protocols and system reliability.

4. Advantages and Weaknesses of Model Agnostic Techniques

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

LIME is celebrated for its flexibility, allowing application across any machine learning model, which is particularly advantageous in personalized medicine for interpreting patient-specific predictions. However, LIME primarily focuses on local explanations, which may not represent the model's behavior more broadly, potentially misleading users about general decision-making processes. Furthermore, LIME's explanations can vary with minor changes in input data, a feature that can undermine the reliability of its interpretations [7].

When to Use LIME:

1. When a simple, interpretable explanation is needed for individual predictions.
2. When the model's global structure is complex but local behavior is linear.
3. When computational resources are limited, as LIME can be more efficient for local explanations.

4.2 *SHapley Additive exPlanations (SHAP)*

SHAP excels by providing consistent and fair contributions of each feature to predictions, rooted in Shapley values from cooperative game theory, making it extremely useful in sectors like finance for detailed risk assessment. However, its application can be computationally intensive, particularly with models having numerous features, which poses challenges in real-time applications. Additionally, the richness of the data SHAP provides can be overwhelming and may require specialized knowledge to interpret effectively [8].

When to Use SHAP:

1. When a globally consistent and theoretically sound explanation is required.
2. When dealing with high-stakes applications where accuracy and reliability of explanations are critical.
3. When the computational cost can be managed, possibly with approximations like TreeSHAP for specific model types (e.g., tree-based models).

4.3 *Partial Dependence Plots (PDP)*

PDPs are praised for their straightforward visualization of the impact of features on predictions, offering clarity that is invaluable in strategic decision-making processes such as in retail pricing strategies. However, PDPs do not account for interactions between features, potentially leading to misleading conclusions where such interactions are significant. They also become less practical with the increase in the number of features, suffering from the curse of dimensionality [9].

4.4 *Individual Conditional Expectation (ICE) Plots*

ICE Plots provide detailed insights at the instance level, showing how predictions change with variations in features for individual data points, which is useful in diverse applications such as real estate valuation. However, the plots can become cluttered and difficult to interpret with large datasets, and like PDPs, they struggle with scalability when dealing with numerous features [10].

4.5 *Anchors*

Anchors offer high precision in their explanations by defining specific rules under which predictions remain invariant, making them robust against variations in input data. This precision is particularly beneficial in critical applications like autonomous driving. Nonetheless, Anchors typically provide narrow insights focused on specific conditions and may not give a comprehensive understanding of the model's overall behavior. The complexity of the rules can also vary, with more complex rules being harder to understand and apply [11].

4.6 *Comparative Analysis*

Our comparative analysis of model-agnostic techniques highlights their varying strengths and weaknesses across different dimensions. To provide a holistic view, we present a summary table (Table 1) capturing key characteristics and performance metrics of each technique.

Table 1. Comparative Analysis of Model-Agnostic Techniques

Technique	Interpretative Clarity	Computational Efficiency	Scalability	User-Friendliness	Application Examples
LIME	High	Moderate	Moderate	High	Medical diagnosis, Customer churn analysis
SHAP	Very High	Low	Low	Moderate	Loan approvals, Risk assessments
PDP	Moderate	High	High	Very High	Sales forecasting, Pricing strategies
ICE	High	High	High	Moderate	Property valuations, Personalization
Anchors	Very High	Low	Moderate	High	Autonomous driving, Safety systems

5. Future Directions

The field of Explainable Artificial Intelligence (XAI) is evolving rapidly, with model-agnostic techniques playing a crucial role in advancing transparency and trust in AI systems across various industries. Looking forward, several key areas hold promise for further research and development:

5.1 *Integration of Model Agnostic Explanations into AI Development Cycles*

Future research should focus on integrating explainability seamlessly into the AI development process, not just as a post-hoc analysis tool. This integration can ensure that AI models are inherently more interpretable and trustworthy from the ground up. Practical frameworks and tools that embed model-agnostic explanations into the model training process could enhance the usability and effectiveness of AI applications in real-world scenarios.

5.2 *Improving Computational Efficiency*

Given the computational demands of techniques like SHAP, especially in large-scale applications, there is a significant need to develop more efficient algorithms that can provide quick and accurate explanations. Research into approximation algorithms or the development of hardware better suited to these tasks could mitigate current limitations.

5.3 *Enhancing Global Interpretability*

While model-agnostic methods excel in local explanations, their ability to provide global insights is often limited. Future efforts could focus on developing methods that offer a clearer picture of overall model behavior while maintaining the flexibility and robustness of model-agnostic approaches. This could involve hybrid techniques that combine the strengths of both local and global explanatory methods.

5.4 *Standardization of Explanation Metrics*

The field would benefit from standardized metrics to evaluate the quality and utility of explanations. Such metrics would enable more objective comparisons between different XAI techniques and facilitate the development of industry standards for explainable AI. Research into what constitutes an effective explanation from both a technical and human-centered perspective is needed.

5.5 *Addressing Ethical and Legal Implications*

As XAI continues to grow, so too will its ethical and legal implications, particularly in sectors like healthcare and finance where decisions can have profound impacts. Future research should explore the ethical dimensions of explanations, ensuring that they are not only effective but also fair and non-discriminatory. Additionally, legal scholars and technologists need to collaborate to ensure that explainability standards meet evolving regulatory requirements.

5.6 *Cross-Disciplinary Approaches to Explainability*

The complexity of explainability challenges warrants a cross-disciplinary approach that incorporates insights from psychology, cognitive science, and legal studies. Understanding how different stakeholders perceive and use AI explanations can inform the design of more effective and user-friendly explanatory tools.

6. Conclusions

This study has conducted a comprehensive comparative analysis of five prominent model agnostic techniques for explainable artificial intelligence: LIME, SHAP, PDP, ICE, and Anchors. Each technique offers unique advantages and faces specific challenges in facilitating the interpretability of AI models across various industries. From healthcare to finance and automotive to retail, the application of these techniques demonstrates a critical balance between computational efficiency, user friendliness, and the depth of explanation required.

Our findings reveal that techniques like SHAP and LIME provide detailed insights into individual predictions, making them particularly suitable for high-stakes environments where precise, granular explanations are necessary. However, these techniques also require considerable computational resources and present challenges in terms of scalability and ease of interpretation. In our research review, we found that utilizing techniques like SHAP and LIME often requires substantial computational resources. Specifically, high-performance CPUs and GPUs, with at least 64 GB of RAM, are recommended to handle the computationally intensive nature of these techniques. For scalability, particularly in large-scale or real-time applications, it is advisable to leverage distributed computing environments or cloud-based solutions. This setup ensures that the processing demands of these sophisticated XAI techniques are met efficiently, allowing for the timely generation of explanations without compromising performance. On the other hand, PDP and ICE, while less computationally intensive, offer more generalized insights that are easier for non-experts to understand but may overlook important interactions between features.

The study highlights the importance of selecting the right XAI technique based on specific industry needs and constraints. For instance, in sectors where decisions have significant personal or financial implications, such as healthcare and finance, the depth and fidelity of the explanation provided by SHAP and LIME are invaluable. Conversely, in sectors like retail, where strategic decisions often rely on broader data trends, PDP and ICE plots provide adequate insights with less complexity.

Looking forward, the field of explainable AI faces several critical challenges. These include improving the computational efficiency of complex techniques, enhancing the global interpretability of model behaviors, and developing standardized metrics for evaluating explanation quality. Moreover, as AI systems become increasingly prevalent across various sectors, the ethical and legal implications of their decisions will necessitate further rigorous research and cross-disciplinary collaboration to ensure that AI remains both innovative and accountable.

In conclusion, this analysis underscores the necessity of continued development in explainable AI technologies. As AI continues to evolve, so too must the techniques that allow humans to understand and trust these advanced systems. By advancing research in model agnostic explainable AI, we can help ensure that AI technologies are used responsibly, ethically, and effectively across all sectors of society.

Conflict of interest

There is no conflict of interest for this study

References

[1] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.

[2] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, p. e1312, 2019.

[3] J. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations," *N. Engl. J. Med.*, vol. 376, no. 26, pp. 2507–2509, 2018.

[4] B. Kim, "Interpretable Machine Learning in Autonomous Driving: Understanding the Decisions of Neural Networks," *Neural Netw.*, vol. 134, pp. 105–115, 2021.

[5] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[6] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. 30 (NIPS 2017)*, Long Beach, CA, USA, Dec. 4–9, 2017.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, Aug. 13–17, 2016.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. 30 (NIPS 2017)*, Long Beach, CA, USA, Dec. 4–9, 2017.

[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[10] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Stat.*, vol. 24, no. 1, pp. 44–65, 2015.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2–7, 2018.