IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 2, May 2025

Impact Factor: 7.67

Enhancing Model Interpretability: A Study on Explainable Machine Learning Techniques

Anil Kumar Chikatimarla[1], Katyayini Gona[2], Teja Sri Oleti[3]

Head, Department of Computer Science[1]

Lecturer, Department of Computer Science[2,3]

A.G. & S.G. Siddhartha Degree College of Arts & Science, Vuyyuru, Andhra Pradesh, India

Abstract: *The importance of machine learning models that are simple for humans to comprehend and apply has grown in fields such as healthcare, autonomous systems, and finance, where people's lives are at stake. Despite these models' effectiveness, user trust and regulatory compliance will be significantly diminished due to their opaque nature. Four state-of-the-art XAI methods—LIME, SHAP, PDP, and CFLE—are compared in this article. We evaluate the algorithms based on how well they express their computations in terms of clarity, efficiency, and accuracy. A group of data scientists and domain experts conducted a user-centered assessment to test these approaches in an actual setting.Evidence from the findings shows that SHAP is accurate, but its computing cost is too high for real-time jobs. The complexity of the ideas is best shown by LIME, even if PDPs and Counterfactual Explanations seem simple at first glance. There is no silver bullet since accuracy and clarity are not mutually exclusive. While considering how to make XAI approaches more relevant, it is crucial to consider potential future research objectives for the field. Here you may discover hybrid explainability approaches, area-specific evaluations, and explanations that happen in real time. This study aims to analyze the current and future of explainability approaches to make machine learning more interpretable.*

Keywords: Explainable Artificial Intelligence (XAI), Machine Learning Interpretability, LIME, SHAP, Model Transparency, Trust in AI

I. INTRODUCTION

A growing number of large companies are showing interest in machine learning. Some examples of places that employ such systems are automated cars, banks, hospitals, and the military. People may not fully grasp the predictive potential of more complex models like ensemble techniques, ANNs, and SVMs. This enigma, often called a "black box problem," will have far-reaching consequences for businesses, universities, and society.Failure to use lessons learned from the past to guide our behavior in the present and future increases the likelihood of high-stakes situations, unexpected events, and trust issues. It is adequately safeguarded by the General Data Protection Regulation (GDPR) and other EU regulations that individuals have a "right to explanation."

Rapid growth around explainable AI has as its primary objective the simplification of machine learning model comprehension, prediction, and evaluation. Improved model debugging, audits of fairness, and regulatory compliance are all made feasible by better interpretability, which boosts users' confidence. Due to the inherent trade-off between model complexity and openness, explainability tools are difficult to design, regardless of their importance. Due to their better efficiency, user-friendliness, and explanatory clarity, deep neural networks are displacing decision trees and other basic models. We focused on building post-hoc explainability approaches to comprehend complicated models without changing their structure.

This article examines and contrasts several XAI approaches based on expert views and empirical evaluations. Our emphasis is on technical specifications and human-centered design to develop ML systems that are both practical and easy to understand and use for addressing problems.

DOI: 10.48175/IJARSCT-26225

II. LITERATURE REVIEW

2.1 Evolution of Interpretability in Machine Learning

Initially, machine learning models that had intrinsic interpretability included decision trees, Naïve Bayes classifiers, and linear regression. The experts agree that purchasers should search for signs of high-quality inputs and solid prediction abilities when assessing algorithms [1]. More effort is required to provide accurate and understandable projections. Such techniques include, for example, random forests [2] and deep neural networks [3].

2.2 Need for Explainable AI

There are a few reasons why interpretability is crucial.

- When people trust each other, they are more likely to embrace technology that is easy to use [4].
- Legislation such as the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act aims to ensure that people can comprehend the reasoning behind AI-related choices that might affect them.
- When it comes to verification and debugging, more training on the models that detect biases, errors, and vulnerabilities might be useful.
- The FAT principles are fundamental to ethical AI and should be followed by all open-source initiatives. Maintaining this code of conduct requires being truthful, frank, and accountable.

2.3 Categories of Explainability Techniques

2.3.1 Model-Specific Interpretability

- Apply a series of logical criteria using decision trees to arrive at a result.
- To make sense of certain models, this is required.
- Classification rules simplify include if-then statements when developing rule-based models.
- For prediction purposes, these models may not be able to handle complex, large-scale datasets with several dimensions.

2.3.2 Post-hoc Model-Agnostic Methods

Post hoc processes must be established to attain the objectives of clarity and predictability:

Known as LIME, it was published by Ribeiro et al. [6]. Using this method, a human-friendly surrogate model may be trained to mimic the behavior of any classifier that is near a certain prediction.

To enhance consistency and geographical accuracy, Lundberg and Lee [7] proposed SHAP, an acronym for "SHapley Additive Explanations," as a system to use. Beginning with cooperative game theory as its cornerstone, SHAP employs a wide variety of explaining techniques.

Technique	Strengths	Weaknesses
LIME	Fast, intuitive, flexible	Can be unstable, sensitive to sampling
SHAP	Theoretically grounded, consistent	Computationally expensive

Popular alternative methods include integrated gradients [9], Partial Dependence Plots (PDP) [10], and Counterfactual Explanations [8].

2.4 Challenges in Explainable AI

- Processing large volumes of data and creating complex models put a premium on resources, making scalability crucial.
- An individual's level of cognitive capacity and prior knowledge dictate the quality of a "good" explanation [11].
- Both oversimplified and under sophisticated models may lead to inaccurate representations of real-world behavior, with the former making it harder to understand nuances.
- To hide their prejudices and wrongdoings, dishonest individuals often resort to lying [12].

2.5 Existing Surveys and Gaps

While many studies have surveyed and categorized explainability techniques, few have compared empirical investigations that assess trade-offs across different domains and models [13, 14]. Unfortunately, the perspectives of practitioners about the viability of these procedures have received very little attention.

The purpose of this study is to fill this significant information gap on the use of explainable ML systems by analyzing survey data and conducting experiments.

III. METHODOLOGY

The methodology of this work is centered on a thorough evaluation of several popular Explainable AI (XAI) approaches, with an emphasis on computing efficiency, comprehensibility, and integrity. The evaluation procedure for XAI techniques primarily consists of two parts: first, reviewing the approaches conceptually; and second, applying the theory to real-world datasets.

Following is further information on the methodology, which includes the datasets used, the XAI techniques used, and the user study conducted to assess interpretability.

3.1 Selection of Explainable AI Techniques

Here are a few well-known XAI algorithms:

By using the LIME (Local Interpretable Model-agnostic Explanations) approach to condense the huge, opaque model, a smaller, more interpretable model that is near to a prediction is generated. On a regional scale, this approach may give valid explanations. Because it is not model-specific, LIME's proposed technique may be used by any ML model.

Another name for SAP is Shapley Additions. A Tool for Education. We ensure that no characteristic substantially impacts a model's prediction by using SHAP, a method derived from cooperative game theory. Complex models, such as gradient-boosted trees and random forests, exhibit significantly improved performance because of SHAP's local and global interpretability.

A helpful tool for understanding the relationship between a model's parameters and their output is the Partial Dependence Plot (PDP). The weights given to each outcome variable are shown graphically below.

Theories on potential other outcomes: They could help us understand the model more if they could show us how it predicts outcomes. We hope that by using this approach, we can pinpoint where the model falls short in drawing adequate conclusions.

Dataset Selection

Our chosen XAI algorithms were put to the test on a plethora of benchmark datasets spanning many disciplines to guarantee complexity and variety. Datasets used for this investigation are listed below:

One way to forecast someone's salary, specifically whether they make more than $50,000 a year or less, is to utilize the Adult Income Dataset, which is also known as the Census Income Dataset. This dataset contains demographic and employment-related information.

The passengers' private information is included in the Titanic Dataset. Analysis of demographic variables such as age, sex, socioeconomic status, and embarkation point will reveal the survivors.

The Boston Housing Dataset: a regression assignment that estimates Boston real estate values based on dimensions such as number of bedrooms, neighborhood safety, and proximity to major employers.

3.3 Model Selection

Multiple machine learning models with varying degrees of complexity were trained for each dataset:

With logistic regression, you may compare two sets of data using a straightforward and easily understandable linear model.

The Random Forest Classifier/Regressor is a black-box model that offers excellent accuracy and is an ensemble approach.

DOI: 10.48175/IJARSCT-26225

One boosting technique with even more complicated decision limits is the Gradient Boosting Machine (GBM), which often beats random forests.

SVM, or Support Vector Machine, is a very successful model, especially in areas with a lot of dimensions.

3.4 Experimental Setup

Seemingly, this is the experimental setup:

The first step was to train the models using every available dataset. We began by making performance estimates using cross-validation and the default hyperparameters, since we believed them to be fair.

To make each model's predictions more understandable, we applied the XAI Techniques: LIME, SHAP, PDPs, and Contextual. Title 7: A Comprehensive Review We found that our LIME model outperformed human forecasts.

Find out how important qualities are by finding out their Shapley values.principal component analysis's findings: These charts analyze traits.

When dealing with counterfactuals, it is conceivable to reorganize predictions while making small changes to attributes.

Finally, Evaluation Tools: A model is said to have "fidelity" when there is a high degree of agreement between its predictions and its actual behavior.

Make it easy to use so that any user may access the data. Immediacy of explanation delivery is one indicator of computational efficiency.

3.5 Studies on Customers

Its usefulness was evaluated by user research. People who took part were split into:

Data scientists are masters of analysis techniques pertaining to interpretability and machine learning.

Specialists in the Field: People with a wealth of knowledge but less experience with machine learning may be found in industries such as healthcare, finance, and marketing.

After viewing the following examples, participants were to rate them:

Be sure that ideas may be readily grasped by the reader.

How feasible it is to put into practice effectively.

Reliability: faith in the rationales offered for choices. The comments have highlighted the benefits and drawbacks of several XAI methods.

3.6 Analyzing Statistics

We used paired t-tests on all datasets and models to find out whether the methods differed significantly in terms of computing efficiency, comprehensibility, and fidelity.

IV. EXPERIMENTAL SETUP

4.1 Selecting a Model and Changing Hyper parameters

To accurately evaluate the explainability approaches, we employed a variety of models based on machine learning with varying degrees of complexity and performance. The complexity of these models varies, ranging from straightforward interpretable models to high-performance black-box models.

4.1.1 The tree-based classifier Randomly

Random Forest is an ensemble learning approach that is based on decision trees. It makes use of many overfit-resistant decision trees, each trained on a unique set of attributes. Because of its exceptional performance and restricted interpretability, it is a popular choice.

There are one hundred trees here.

The minimum leaf in the sample is one.

The minimum split is two.

The maximum depth is 10.

DOI: 10.48175/IJARSCT-26225

4.1.2 Classifier Extreme Gradient Booster or XGBoost:

It is a scalable and very efficient gradient boost framework implementation. To optimize performance, it makes use of regularization and parallel processing.

Hyperparameters:

The learning rate is 0% and there are 200 estimators.

Maximum depth: 6;

subsample: 0.8

4.1.3 Neural System Networks:

The comprehension of deep learning models created by XAI techniques was evaluated using the fully connected neural network (FCNN). Despite their strength, neural networks are frequently challenging to comprehend.

Hyperparameters: The number of mystery units is sixty-four.

The function used for activation is ReLU, the batch size is 32, and the training rate is 0.01.

4.1.4 Modification of Hyperparameters

Grid Search and five-fold cross-validation were used to modify the hyperparameters. The best model performance was guaranteed across a range of parameter combinations thanks to this methodical approach.

4.2. Data Collection Methods:

Some interesting recommendations from the XAI group include:

4.2.1 The abbreviation "LIME" stands for "local interpretable model-agnostic explanation."

One potential result of using the LIME technique is discovering a simple model, maybe a linear one, that can replicate the local behavior with few data changes.

Data Necessary to Finish the Assignment:

To train a linear classifier, you may use Lime, a Python program, and 500 tabular data samples.

There is a 0.75 cm breadth to the kernel.

What the Shapley Drug Test Is and Why It Matters

4.2.2 SHAP

To clarify the model's findings, SHAP may adjust the weights of each attribute using the Shapley values from cooperative game theory.

Items Required to Finish the Job:

Complete sets of all required legal papers.

Python offers a wide variety of programs that may solve a wide variety of issues.

Some examples of these programs are kernel form, rough approximation, and many more.

4.2.3 Partial Dependence Plots (PDP)

While considering the average impacts of all features, PDP displays the feature's marginal influence on the model's prediction.

Instructions for Execution:

Sklearn and pdpbox are Python packages.

The top three traits, as determined by the model's relevance, are shown in the grid, which has 20 points.

4.2.4 Counterfactual Explanations

According to counterfactual theories, altering the input qualities just a little bit would be the sole way to alter the model's prediction.

Details of the implementation: the Nearest Neighbor Search method and Euclidean distance measuring

DOI: 10.48175/IJARSCT-26225

There can be no more than ten iterations.

4.3 Performance Metrics

The models and XAI techniques were assessed using the performance metrics listed below:

Fidelity: This measure, which is based on SHAP consistency or surrogate model correctness, indicates how well the explanation captures the behavior of the real model.

Comprehensibility: Evaluated according to user research comments about how useful and clear the explanations are.

Computational efficiency is the amount of time (in seconds per occurrence or global explanation) required to produce an explanation.

V. RESULTS AND DISCUSSION

5.1 Quantitative Evaluation

5.1.1 Performance of Models

Before evaluating the explainability methods, we assessed the predictive performance of the machine learning models across the selected datasets. The classification accuracy and regression scores were as follows:

Model	Adult Income Accuracy (%)	Titanic Survival Accuracy (%)	Boston Housing RMSE
Random Forest	85.2	79.3	4.5
XGBoost	88.1	81.7	4.1
Neural Network	88.6	82.5	3.9

Findings:

Neural networks performed better than other models, particularly on the Boston Housing dataset.

XGBoost outperformed Random Forest in classification challenges.

5.1.2 Explainability Technique Evaluation

Fidelity Scores:

Model	LIME Fidelity (%)	SHAP Fidelity (%)
Random Forest	87.2	91.5
XGBoost	89.3	92.7
Neural Network	85.4	88.9

According to the results, SHAP explanation were more accurate for all models and more consistent with the behavior of the actual models, while LIME explanations were a little less accurate, especially for complicated models like neural networks.

Computational Time:

Technique	Average Computation Time per Instance (seconds)
LIME	1.5
SHAP	3.1
PDP	0.3
Counterfactual	2.8

Results:

PDP finished the calculations the quickest.

While LIME provided a fair compromise among speed of computation and explanation quality, SHAP was highly accurate despite its substantial human costs.

Comprehensibility (User Study Ratings):

Explanation Method	Data Scientists (5)	Domain Experts (5)	Non-Experts (5)
LIME	4.3	4.0	3.5
SHAP	4.5	4.2	3.8
PDP	4.1	4.3	4.2
Counterfactual	4.0	4.5	4.1

Results:

Data scientists selected SHAP because of its accuracy and comprehensiveness, whereas domain experts preferred PDPs and counterfactuals because they were the easiest to comprehend.

PDPs were the simplest for non-experts to comprehend.

5.2 Discussion

5.2.1 Trade-offs Between Fidelity and Comprehensibility

The most accurate answers are given by SHAP, although it can occasionally be too complicated for non-experts.

Though marginally less precise, LIME provides more reliable local responses.

5.2.2 Computational Efficiency vs. Real-Time Use

Its lower processing overhead makes methods such LIME and PDPs more appropriate for applications that operate in real time (such medical diagnostics and fraud detection).

Because of its long processing time, SHAP is not suitable for circumstances in which time is of the importance.

5.2.3 Practitioner Insights

According to the practitioner survey, the most crucial element in adoption was credibility.

Domain-specific specialists, such as medical professionals, stressed that regardless of technical accuracy, any explanation is acceptable as long as it is understandable and beneficial.

Data scientists use SHAP for model debugging and compliance checks.

VI. CONCLUSION

Rapid advances in machine learning have enabled highly accurate prediction models. Sophisticated models' incomprehensibility and complexity continue to be major barriers to their general adoption, especially un high-stakes sectors including healthcare, banking, and autonomous systems. This research provided a thorough comparison of a number of cutting-edge Explainable AI (XAI) methods, including LIME, SHAP, PDP, and Counterfactual Explanations, emphasizing their advantages, disadvantages, and applicability to various machine learning model types.

Our results show that fidelity, readability, and computing efficiency must be carefully balanced:

The most accurate answers are regularly given by SHAP, although it is less suitable for real-time applications due to its higher computing cost.

For deep learning models in particular, LIME strikes an appropriate equilibrium among interpretability and efficiency at the expense of some fidelity; personal development plans and Counterfactual Explanations provide more logical, human-understandable insights, but they are less able to capture intricate model behaviors.

The user survey also found that the value of an explanation is significantly influenced by end-user skill:

Technology practitioners and data experts favored SHAP and LIME because of their comprehensive and precise insights.

Both subject matter specialists and non-technical users found counterfactual explanations and personal growth goals to be more beneficial and understandable.

In the end, it may be claimed that such statistic does not always offer a better reason. The target audience, underlying model complexity, and application domain should all be considered when selecting a XAI strategy. The best approach

DOI: 10.48175/IJARSCT-26225

is probably a hybrid one that incorporates several different explaining techniques. By offering empirical support for trade-offs between significant XAI technologies and useful recommendations for their implementation in real-world contexts, this work advances interpretable machine learning.

VII. FUTURE WORK

This work provides valuable insights into the compromises among fidelity, comprehensibility, and processing effectiveness in explainable artificial intelligence (XAI) techniques, but there is still much to learn. To improve its readability and accuracy, future explainable AI research should concentrate on merging many explanation methodologies, such as SHAP and LIME. Examples of businesses that need quicker real-time explanations include healthcare and self-driving cars.

REFERENCES

[1]. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.

[2]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[3]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

[4]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*, arXiv:1702.08608.

[5]. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation.” *AI Magazine*, 38(3), 50–57.

[6]. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

[7]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

[8]. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

[9]. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning (ICML)*.

[10]. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2).

[11]. Lipton, Z.C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.

[12]. Slack, D., Hilgard, S., Jia, E., Singh, S., &Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[13]. Adadi, A., &Berrada, M. (2018). Peeking inside the black-box: A survey of explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.

[14]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., &Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1–42