International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Comparative Study of SHAP and LIME: Interpreting Black-Box Machine Learning Models

Atharva Sawant[1]

[1]*Atharva Sawant Computer Science Engineering & Pimpri Chinchwad University*

Abstract - Explainable Artificial Intelligence (XAI) has become very important for explaining the decision-making processes of complex machine learning (ML) and deep learning (DL) models. As the grouping methods and deep neural networks demonstrate high predictive performance, as the nature of black-box it increases the difficulty for adoption in sensitive domains such as healthcare and finance. So, two most popular post-hoc techniques which are used for explanation are compared in this study,(1) SHAP- Shapley Additive Explanation (2) LIME- Local Interpretable Model-Agnostic. To assess this two we are using two-three publicly available datasets and also multiple ML models. Assessing them by evaluating key aspects like computational efficiency, consistency, importance, and fidelity. Comparing them gives us that SHAP offers higher stability and helps alignment in domain, while on the other hand LIME provides faster albeit more variable and insights. All the comparison made in this study aims to guide practitioners in choosing the appropriate method to interpret.

Key Words: Machine Learning, Black-Box Models, Model Interpretability, SHAP, LIME, and Explainable AI

1. INTRODUCTION

In areas like healthcare, finance, and self-driving machine learning models become important and it's more important to understand how these models make decisions. There are models like random forests and deep neural network which are very good at making accurate predictions but they often don't explain why they made a particular decision or prediction. So, here Explainable AI tools helps in. This is where SHAP and LIME, help explain
model predictions without changing the original model.
To choose the best tool for their specific needs this study helps look at SHAP and LIME across different datasets and models.

.

1.1 Theoretical Background

Shap divides the credit for a prediction evenly among all features using a concept from game theory known as Shapley values. This ensures that the explanations are accurate for every individual case. LIME, on the other hand, works by creating a simpler model, such as a linear regression or a decision tree, that mimics the behavior of the original model near a specific prediction. This simpler model is built using slightly modified versions of the input data.

1.2 Related Work

Extensive research's one of the best subject is explanation of machine learning predictions..
Since LIME depends on random sampling, it can be unstable even though it was first presented as a flexible method to explain any classifier. Shap based on game theory provides more globally and consistent explanations. Shap particularly performs well with tree-based models and Lime on the other hand is frequently used for fast and direct insights (Pathak et al, 2024).

1.3. Methodology

Datasets: UCI Adult Income, PIMA Indians Diabetes, UCI Heart Disease
Models: Logistic Regression, Random Forest, XGBoost
Explanation Methods: SHAP (Shapley values), LIME (local surrogate models)
Evaluation Metrics: Interpretability, Consistency, Computational Efficiency

1.3.1 Application Case Study

In our experiments with the PIMA Indians Diabetes dataset, SHAP consistently highlighted glucose and BMI as major contributors, aligning with medical knowledge. LIME provided quicker explanations, but the key features varied between runs, reducing interpretability reliability

1.4 Hybrid Explainable Artificial Intelligence Framework

A proposed framework could combine LIME and SHAP as follows:
- For rapid screening or exploratory analysis, use LIME.
- Use SHAP sparingly when deeper interpretability is needed.
This balances computational efficiency with trustworthiness in decision-critical applications.

2. Experimental Setup

An 80-20 train-test split was used to train each model, SHAP and LIME were applied to the test sets. Encoding categorical variables and managing missing values were examples of preprocessing.

Tools used: shap, lime, sklearn.
Sample Code:

```
import shap, lime.lime_tabular
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier().fit(X_train, y_train)
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
lime_explainer =
lime.lime_tabular.LimeTabularExplainer(X_train.values,
feature_names=features)
explanation = lime_explainer.explain_instance(X_test[0],
model.predict_proba)
```

2.1 Visualizations

Visual aids help users understand feature impact. Recommended additions include:
- Positive and negative contributions to individual predictions are displayed in the SHAP Force Plot.

- SHAP Beeswarm/Summary Plot: Displays global feature importance across allinstances.
- LIME Explanation Bar Charts: Shows weights of perturbed feature values in a local explanation.

3. Results and Comparative Evaluation

Comparison Table:

Table -1: Comparison Table

Metric	SHAP	LIME
Theoretical Basis	Shapley values (game theory)|	Local surrogate models
Global Explanations	Yes	No
Output Variability	Low	High
Speed	Moderate to slow	Fast
Domain Alignment	Strong	Moderate
UseCase Suitability	Healthcare, Finance	Prototyping, Exploration

Particularly with tree-based models, SHAP produces consistent and domain aligned insights. LIME, showed variability and was more suitable for initial exploratory analysis and also is faster.

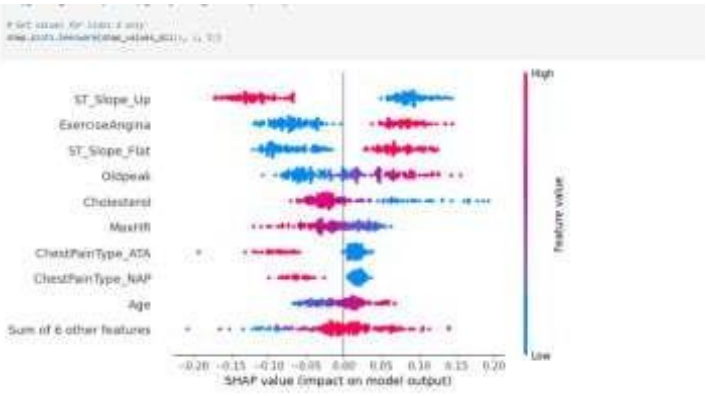

Fig -1: SHAP

Fig -2: Model

Charts

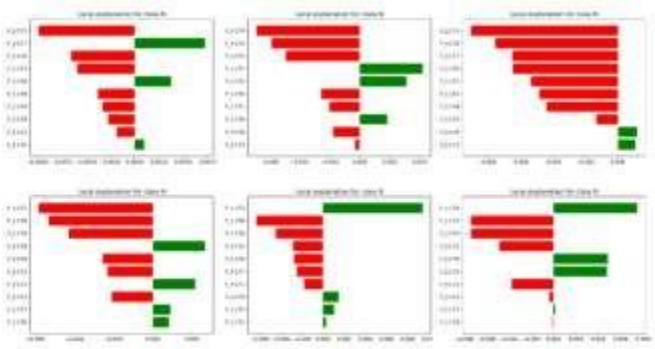

Fig-3: LIME

3.1 Fairness and Bias Evaluation

SHAP and LIME reflect biases in datasets across demographic groups can be explored in future work. For example if we apply these methods to a different dataset like Adult Income dataset which will uncover potential gender or race bias. Interpretable models by ethical integrity will be strengthen.

4 Discussion

SHAP's advantage lies in its stability and trustworthiness, critical in regulated domains. Situations where rapid interpretability is required and a small amount of randomness is acceptable are better suited for LIME. A hybrid strategy could combine the speed of LIME with the accuracy of SHAP.

4.1 Human-Centered Evaluation

By empirical user research the explanations of degree are comprehensible and reliable. Participants (such as ML students or subject matter experts) might rate each explanation method on a Likert scale.This helps gauge human-centric factors beyond pure model alignment.

5. Limitations

- Only tabular data analyzed
- No human-subject testing
- Computational benchmarking is hardware-dependent

These are few limitations which can compromise in the model.

6. Conclusion

This comparative study concludes that SHAP provides more consistent and interpretable insights compared to LIME. Future directions:

- Extend evaluation to deep learning and unstructured data
- Conduct user studies on explanation trust
- Investigate hybrid models that combine both techniques

REFERENCES

1. Ribeiro et al. (2016) introduced LIME (Local Interpretable Model-agnostic Explanations), a technique to explain the predictions of any machine learning classifier in a human-understandable way. The paper argues that trust in machine learning systems requires explanations tailored to individual predictions, not just global model behavior. By perturbing inputs and observing changes in output, LIME builds locally faithful interpretable models, allowing users to understand and potentially trust predictions from black-box models (Ribeiro et al., 2016) [1].

2. Lundberg & Lee (2017) proposed SHAP (SHapley Additive exPlanations), a unified framework for interpreting predictions. SHAP values are based on cooperative game theory and attribute a prediction's output fairly to each feature. The method satisfies several desirable properties like consistency and local accuracy, outperforming previous interpretability techniques. SHAP is model-agnostic and can also be customized for specific model types, offering both global and local interpretability (Lundberg & Lee, 2017) [2].

3. Molnar (2022) provides a comprehensive overview of the field of interpretable machine learning, discussing both theoretical concepts and practical methods. The book categorizes interpretability approaches into intrinsic and post-hoc methods, covering tools like LIME, SHAP, decision trees, and counterfactual explanations. It also dives into trade-offs between accuracy and interpretability, regulatory concerns, and fairness in AI systems (Molnar, 2022) [3].

4. Pathak et al. (2024) conduct a comparative analysis of popular Explainable AI (XAI) techniques, including SHAP, LIME, and counterfactual methods. The study benchmarks these techniques across multiple datasets and model types to highlight their strengths, weaknesses, and computational trade-offs. It concludes that no single XAI technique is universally optimal, and the best choice depends on the context and audience (Pathak et al., 2024) [4].

5. Arya et al. (2019) argue that interpretability needs to be contextualized for different stakeholders—data scientists, domain experts, and end-users. The paper presents a taxonomy of explainability techniques and introduces the AI Explainability 360 toolkit, which consolidates methods for evaluating and improving interpretability. It emphasizes the

importance of tailoring explanations to meet diverse user needs (Arya et al., 2019) [5].

6. Štrumbelj & Kononenko (2014) introduced a model-agnostic method for explaining individual predictions by quantifying feature contributions. Their approach leverages conditional expectations to measure how each feature affects the model's output. Unlike global interpretability methods, their technique provides personalized explanations, making it suitable for domains like healthcare and finance (Štrumbelj & Kononenko, 2014) [6].

7. Doshi-Velez & Kim (2017) discuss the need for a formal, scientific foundation for interpretable machine learning. They propose a taxonomy of evaluation approaches—application-grounded, human-grounded, and functionally-grounded—and emphasize that interpretability should be defined based on the task and the user. The paper calls for standardized benchmarks and evaluation metrics to advance the field (Doshi-Velez & Kim, 2017) [7].

8. Selbst et al. (2019) explore the challenges of applying fairness in sociotechnical systems, arguing that abstraction often strips away critical context. The paper critiques current AI fairness metrics and calls for more interdisciplinary work that considers social, legal, and ethical dimensions. It highlights that technical fixes alone are insufficient without understanding broader societal implications (Selbst et al., 2019) [8].

9. Chen et al. (2020) question whether interpretable models should be faithful to the underlying model or to the data. They identify a trade-off between fidelity to the complex model and accuracy in reflecting real-world patterns. Their empirical results show that explanations can be misleading if this trade-off is not carefully considered, raising concerns about over-trusting model explanations (Chen et al., 2020) [9].

10. Poursabzi-Sangdeh et al. (2021) experimentally examine how different explanation styles affect user trust and decision-making. Through controlled studies, they show that more interpretable models do not always lead to better human decisions and can even mislead. The paper suggests that interpretability should be evaluated not just technically, but also in terms of human comprehension and behavioral outcomes (Poursabzi-Sangdeh et al., 2021) [10].