



An Analysis of Explainable AI Techniques in Machine Learning

Author: **Riddhi Mitul Brahmbhatt**

Program: MSc Data Science

College: Kandivali Education Society's BK Shroff College of Arts and MH Shroff College of Commerce
(Autonomous)

riddhibrahmbhatt6@gmail.com

How to Cite this Article:

Brahmbhatt, R. M. (2026). An Analysis of Explainable AI Techniques in Machine Learning. International Journal of Creative and Open Research in Engineering and Management, 2(10), 1-9.
<https://doi.org/10.55041/ijcope.v2i9.247>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i9.247>

Abstract

This research paper discusses Explainable Artificial Intelligence (XAI) techniques used in Machine Learning and their importance in understanding machine learning predictions. Machine learning models are increasingly used in different areas such as healthcare, finance, education, customer analytics and business decision-making. However, some advanced machine learning models can produce predictions without clearly showing how or why a particular decision was made. This creates a challenge because users may find it difficult to understand, verify or trust the output of the model.

This research focuses on exploring and understanding Explainable AI, different types of explanation techniques, their uses, challenges and possible solutions. It discusses commonly used techniques such as Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), permutation feature importance and surrogate models. XAI techniques can help users understand the contribution of different features to machine learning predictions. SHAP, for example, provides feature contribution explanations, while LIME explains individual predictions by approximating a complex model locally.

The study follows descriptive research based on secondary data collected from research papers, journals, academic publications and reliable online sources. The research aims to provide a better understanding of how XAI techniques can improve transparency, interpretability and understanding of machine learning models while also identifying their limitations and challenges.

Keywords: Explainable AI, XAI, Machine Learning, Interpretability, LIME, SHAP, Model Explainability, Artificial Intelligence

INTRODUCTION

Currently, Artificial Intelligence (AI) and Machine Learning (ML) are being widely used to make predictions and support decision-making. Machine learning models can learn patterns from data and generate predictions without being explicitly programmed for every individual decision. However, some complex models can be difficult for humans to understand because the reasoning behind their predictions is not always visible.

This problem is commonly associated with the idea of a "black-box" model. A black-box model can provide an output or prediction, but users may not easily understand why that output was



produced. As machine learning is increasingly applied to important areas, understanding the reasoning behind predictions has become important.

Explainable Artificial Intelligence, commonly known as XAI, focuses on providing understandable information about how AI or machine learning systems produce their outputs. XAI can help users understand the relationship between input features and model predictions. NIST describes explainability as an important characteristic of trustworthy AI and distinguishes it from interpretability while recognizing that both support understanding of AI systems.

Different XAI techniques have been developed to explain machine learning predictions. Some techniques explain individual predictions, while others provide an overall understanding of how a model behaves. LIME explains individual predictions by creating a simpler local approximation of a complex model, whereas SHAP uses feature contributions to explain predictions. Other techniques include Partial Dependence Plots, permutation feature importance and surrogate models.

Explainability is important because simply producing an accurate prediction may not always be sufficient. Users may also want to understand why a model produced a particular result. For example, if a machine learning model predicts that a customer is likely to leave a service, the user may want to know which factors contributed to that prediction.

Therefore, this research focuses on understanding Explainable AI techniques in machine learning, their purpose, different methods, applications, challenges and possible solutions.

Literature Review

Explainable Artificial Intelligence has become an important research area because of the increasing use of machine learning models in decision-making systems. Complex models may achieve useful predictive performance while making their internal decision process difficult for users to understand. XAI methods attempt to provide explanations that help users understand model outputs.

Research on XAI includes a wide range of methods. A review of XAI methods discusses approaches such as LIME, SHAP, Anchors, Partial Dependence, Integrated Gradients, surrogate models and other explanation techniques.

LIME, or Local Interpretable Model-Agnostic Explanations, is a model-agnostic method that focuses on explaining individual predictions. It creates a simplified model around a particular instance to approximate the behaviour of a more complex model. This makes it possible to understand which input features contributed to a particular prediction.

SHAP, or SHapley Additive exPlanations, is another widely used explanation technique. It uses the concept of Shapley values from cooperative game theory to estimate the contribution of individual features to a prediction. Different SHAP implementations can be used for different types of machine learning models.

Partial Dependence Plots (PDP) can be used to study the relationship between one or more features and the model's predicted output. Permutation feature importance can also be used to understand how model performance changes when feature values are randomly permuted.



Surrogate models provide another approach to explain complex systems. An interpretable model can be trained to approximate the behaviour of a more complex model, allowing users to obtain a simpler representation of the model's behaviour.

NIST has also proposed four principles for explainable AI: an AI system should provide evidence or reasons for its outputs, provide explanations that are meaningful to users, ensure the explanation accurately reflects the system's process, and recognize situations where the system reaches its knowledge limits.

Recent research continues to investigate the reliability and limitations of XAI methods. For example, recent work examining LIME discusses issues such as explanation stability and fidelity, showing that producing an explanation does not automatically mean that the explanation is always reliable or equally useful in every situation.

Research Gap

Explainable AI has been widely studied, and many techniques have been developed for interpreting machine learning models. However, different XAI methods provide different types of explanations and may have different strengths and limitations.

Some techniques focus on individual predictions, while others provide a broader understanding of model behaviour. The usefulness of an explanation can also depend on the type of model, dataset, user and application.

There is therefore a need to understand the differences between commonly used XAI techniques and examine their advantages, limitations and appropriate applications.

Another important issue is that an explanation should not only be understandable but should also accurately represent the behaviour of the model. NIST's XAI principles specifically highlight meaningfulness and accuracy of explanations as important considerations.

Therefore, this research focuses on understanding commonly used XAI techniques, their purposes, benefits, challenges and the situations in which they can be applied.

Areas to Focus

This research focuses on the following main areas:

1. Awareness and Understanding

Understanding what Explainable AI means and why explainability is important in machine learning.

2. XAI Techniques

Understanding commonly used techniques such as:

- LIME
- SHAP



- Partial Dependence Plots
- Permutation Feature Importance
- Surrogate Models

3. Applications

Understanding how XAI can be used in areas such as:

- Healthcare
 - Finance
- Customer analytics
- Education
- Business decision-making

4. Challenges and Limitations

Understanding problems related to:

- Complexity of explanations
 - Stability
- Accuracy of explanations
- Model dependency
- Computational cost
- User understanding

5. Trust and Responsible AI

Understanding how explainability can support transparency and responsible use of machine learning systems.

Research Objective

The main purpose of this research is to analyze Explainable AI techniques used in machine learning and understand their importance in explaining model predictions.

The study aims to achieve the following objectives:

1. To understand the concept and importance of Explainable Artificial Intelligence in Machine Learning.
2. To identify and describe commonly used Explainable AI techniques.
3. To understand the differences between local and global explanation approaches.
4. To study the applications of XAI techniques in different machine learning domains.



5. To identify the major challenges and limitations associated with Explainable AI.
6. To understand how explainability can improve transparency and understanding of machine learning predictions.
7. To contribute to the understanding of suitable XAI techniques for different machine learning applications.

Research Methodology

This study follows **descriptive research** to examine Explainable AI techniques in Machine Learning and is completely based on **secondary data** collected from research papers, academic journals, research publications, official reports and reliable online sources.

The collected information is used to understand the concept of Explainable AI, different explanation techniques, their applications, advantages, challenges and limitations.

The study reviews different XAI methods used to explain machine learning predictions. The techniques are described and compared based on their purpose, type of explanation, model dependency and application.

The research does not claim to experimentally determine which XAI technique is universally superior. Instead, it provides a descriptive understanding of the available techniques and their characteristics.

Types/Methods of Explainable AI

The following techniques are commonly used for explaining machine learning models.

1. LIME

LIME stands for **Local Interpretable Model-Agnostic Explanations**.

LIME is used to explain individual predictions of complex machine learning models. It creates a simpler interpretable approximation around a particular prediction.

For example:

Customer Data



ML Model



Prediction: Customer may Churn

↓ LIME





Important factors for this prediction

LIME is called model-agnostic because it can be applied to different types of machine learning models. However, research has identified challenges involving the stability and fidelity of LIME explanations.

2. SHAP

SHAP stands for **SHapley Additive exPlanations**.

SHAP explains how individual features contribute to a model prediction. It is based on Shapley values from cooperative game theory.

For example, a customer churn model may produce: Prediction: High Churn

Risk

Feature	Contribution

Contract Type	High Monthly Charges
	Medium Tenure High
Payment Method	Low

This helps users understand which features contributed to the prediction.

SHAP can provide local explanations for individual observations and can also be used to understand feature importance across many observations.

3. Partial Dependence Plot

A **Partial Dependence Plot (PDP)** helps show the relationship between a feature and the model's predicted outcome.

For example:

Monthly Charges



Model Prediction



Probability of Churn

The plot can help users understand how changes in a feature are associated with changes in the model's prediction.



PDP is one of the established approaches used for understanding relationships between features and model outputs.

4. Permutation Feature Importance

Permutation feature importance measures the effect of a feature by randomly changing or shuffling its values and observing the resulting change in model performance.

If changing a particular feature causes a large reduction in model performance, that feature may be important to the model's predictions.

For example:

Original Model Performance



Shuffle Feature A



Performance decreases significantly



Feature A may be important

This provides a way to understand the importance of features based on their effect on model performance.

5. Surrogate Models

A surrogate model uses a simpler and more interpretable model to approximate the behaviour of a complex machine learning model.

For example:

Complex ML Model



Prediction Behaviour



Surrogate Model



Simple Explanation

Decision trees can sometimes be used as interpretable surrogate models to provide a simplified representation of a complex model's behaviour.



Problems

1. Complexity of Explanations

Some explanations may be difficult for ordinary users to understand, especially when technical concepts or mathematical information are presented.

An explanation should be appropriate for its intended audience. NIST notes that meaningfulness can depend on the user's role and knowledge.

2. Stability of Explanations

Some XAI methods may produce different explanations when inputs or conditions change slightly.

This creates a concern about whether users can consistently rely on an explanation.

Research on LIME, for example, identifies stability as one of the important challenges in using local explanations.

3. Accuracy of Explanations

An explanation should accurately represent the model's behaviour.

A simple explanation may be easy to understand but may not fully represent the underlying complex model.

NIST therefore identifies explanation accuracy as one of its fundamental XAI principles.

4. Model Dependency

Some explanation techniques work better with particular types of machine learning models.

For example, different SHAP implementations are designed for different model structures, while some methods are model-agnostic.

5. Computational Cost

Some explanation methods can require additional computation, especially when explanations need to evaluate many possible feature combinations or model behaviours.

SHAP implementations, for example, can differ significantly in computational requirements depending on the approach used.

6. Knowledge Limitations

An AI system may produce an output even when it does not have sufficient information or confidence to provide a reliable answer.

Recognizing and communicating knowledge limits is therefore an important part of explainable AI.

7. Difference Between Users

A data scientist, manager and ordinary user may require different levels of explanation.

A highly technical explanation may be useful to a data scientist but difficult for a non-technical user to understand.

Therefore, explanations should be designed according to the needs of the intended user.



Solutions

1. User-Friendly Explanations

XAI systems should provide explanations that users can understand without requiring unnecessary technical knowledge.

2. Selecting Suitable XAI Techniques

Different techniques should be selected according to:

- Model type
- Dataset
- Application
- User requirements
- Type of explanation required

3. Validate Explanations

Explanations should be checked to determine whether they accurately represent model behaviour.

4. Use Multiple Explanation Methods

Using more than one XAI technique can provide different perspectives on a model's behaviour.

For example: SHAP

+ LIME

+

Feature Importance

+ PDP

↓

Better Understanding

However, multiple explanations should be interpreted carefully rather than automatically assuming that agreement means correctness.

5. Explain Knowledge Limits

AI systems should identify situations where their predictions may be unreliable or outside the intended scope of the system. This is specifically emphasized in NIST's knowledge-limit principle.

6. Provide Appropriate Training

Users should receive appropriate guidance about what an XAI explanation means and what it does not mean.

7. Responsible Implementation

Explainability should be considered together with other responsible AI requirements such as reliability, privacy, security and fairness rather than being treated as the only measure of a trustworthy system. NIST's AI risk-management resources treat explainability and interpretability as part of broader AI trustworthiness considerations.



Discussion

Explainable Artificial Intelligence is an important area of Machine Learning because users increasingly need to understand how AI systems reach their predictions. The use of complex machine learning models can make it difficult to identify the factors responsible for a particular prediction.

The research focuses on understanding different XAI techniques such as LIME, SHAP, Partial Dependence Plots, permutation feature importance and surrogate models. These methods provide different forms of explanation and can be applied according to the model, dataset and user requirements.

XAI is not only about producing an explanation but also about producing an explanation that is meaningful and appropriately represents the model's behaviour. NIST's principles emphasize evidence or reasons, meaningfulness, explanation accuracy and recognition of knowledge limits.

Different XAI techniques also have limitations. LIME can face concerns about stability and fidelity, while SHAP can involve computational considerations depending on the implementation and model.

Therefore, Explainable AI should not be considered as one single technique. It is a collection of approaches that can help different users understand machine learning systems from different perspectives.

XAI can be particularly useful in applications where understanding predictions is important. For example, in healthcare, users may want to understand factors contributing to a prediction. In finance, organizations may need explanations for automated decisions. In customer analytics, businesses may want to understand factors associated with customer behaviour.

The purpose of explainability is therefore not simply to make a machine learning model appear transparent. Instead, explanations should help users understand model behaviour while recognizing the limitations of the explanation itself.

Conclusion

In conclusion, Explainable Artificial Intelligence is an important area of Machine Learning that focuses on helping users understand AI and machine learning predictions. As machine learning models become more complex and are used in different decision-making applications, the ability to understand their predictions becomes increasingly important.

This research analyzed different XAI techniques including LIME, SHAP, Partial Dependence Plots, permutation feature importance and surrogate models. Each technique provides a different approach for understanding machine learning behaviour.

The research also identified challenges such as explanation complexity, stability, accuracy, computational requirements, model dependency and knowledge limitations. Therefore, selecting an XAI method should depend on the machine learning model, application, user and type of explanation required.

Explainability can contribute to greater transparency and understanding of machine learning systems, but an explanation should not automatically be treated as proof that a model is correct or fair. Proper evaluation and responsible use are important.

Overall, Explainable AI can support the development and use of machine learning systems where understanding, transparency and responsible decision-making are important.

Continued research can focus on developing more reliable, understandable and user-specific explanation methods.



References

- Phillips, P. J., C. A. Hahn, P. C. Fontana, A. N. Yates, K. K. Greene, D. A. Broniatowski, and M. A. Przybocki. **“Four Principles of Explainable Artificial Intelligence.”** National Institute of Standards and Technology, NISTIR 8312, 2021. <https://doi.org/10.6028/NIST.IR.8312>. Accessed 14 Sept. 2026.
- **“NIST Asks A.I. to Explain Itself.”** National Institute of Standards and Technology, 18 Aug. 2020. <https://www.nist.gov/news-events/news/2020/08/nist-asks-ai-explain-itself>. Accessed 14 Sept. 2026.
- Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. **“Explainable AI Methods – A Brief Overview.”** *xxAI – Beyond Explainable AI*, Lecture Notes in Computer Science, vol. 13200, Springer, 2022, pp. 13–38. https://doi.org/10.1007/978-3-031-04083-2_2. Accessed 14 Sept. 2026.
- Salih, Ahmed M., Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Gloria Menegaz, and Karim Lekadir. **“A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME.”** *Advanced Intelligent Systems*, vol. 7, no. 1, 2025. <https://doi.org/10.1002/aisy.202400304>. Accessed 14 Sept. 2026.
- Shaikh, Abujar S., Rahul M. Samant, Kshitij S. Patil, Nilesh R. Patil, and Aarya R. Mirkale. **“Review on Explainable AI by Using LIME and SHAP Models for Healthcare Domain.”** *International Journal of Computer Applications*, vol. 185, no. 45, 2023, pp. 18–23. <https://doi.org/10.5120/ijca2023923263>. Accessed 14 Sept. 2026.
- **“A Survey on Explainable AI Using Machine Learning Algorithms SHAP and LIME.”** *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024. <https://doi.org/10.1109/ICCCNT61001.2024.10725120>. Accessed 14 Sept. 2026.
- **“Explainable AI Evaluation: A Top-Down Approach for Selecting Optimal Explanations for Black Box Models.”** *Information*, vol. 15, no. 1, 2024, article 4. <https://doi.org/10.3390/info15010004>. Accessed 14 Sept. 2026.