MANTECH Publications
Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 10, Issue 2, July-December 2025

Explainable AI (XAI) & Model Interpretability

Suman Sharma[1], Devansh Prajapati[2], Meera Joshi[3], Jahangir Ali[4]
Associate Professor, Students
Department of Computer Applications
Pragati Degree College, Bhopal, India
Email ID:*Sumansharma12@gmail.com[1], 0devansprajapati@yahoo.com[2], meera_joshi10@rediffmail.com[3]*

ABSTRACT

Artificial Intelligence (AI) systems are increasingly used in critical applications such as healthcare, finance, education, and governance. However, many modern AI models, especially deep learning networks, operate as "black boxes" where decision making process is not easily understood by humans. This lack of transparency leads to problems of trust, accountability, fairness, and regulatory compliance. Explainable AI (XAI) has emerged as an important research area to make AI systems more transparent, interpretable, and trustworthy. This paper presents a comprehensive review of Explainable AI techniques and model interpretability approaches. Various methods including intrinsic interpretability, post-hoc explanations, feature importance, visualization tools, and rule-based explanations are discussed. The paper also highlights the importance of XAI in real world applications and current challenges in the field. Tables and figures are provided for better understanding of different techniques. Finally, future directions of research in XAI are discussed.

KEYWORDS: *Explainable AI, Model Interpretability, Black Box Models, Trustworthy AI, Feature Importance, Post-hoc Explanation, Transparency*

INTRODUCTION

Artificial Intelligence is growing very rapidly in recent years. Many systems are now using machine learning and deep learning algorithms to automate decision making. These systems often achieve very high accuracy but the internal working is complex and not easy to

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

understand. This creates a gap between system performance and human understanding.

Traditional models like linear regression and decision trees were easy to interpret. But advanced models like neural networks, ensemble models, and transformers are difficult to explain. In sensitive domains such as medical diagnosis, loan approval, criminal justice, and autonomous vehicles, it is very important to know *why* a model gives a certain decision.
Explainable AI (XAI) focuses on making AI decisions understandable for humans. It helps to increase trust, detect bias, improve debugging, and satisfy legal regulations. XAI is not only about explanation but also about creating responsible AI systems.

NEED FOR EXPLAINABLE AI (XAI)

As Artificial Intelligence systems become deeply embedded in real-world decision making—such as healthcare diagnosis, financial approvals, autonomous driving, smart surveillance, and IoT-based automation—the demand for **Explainable AI (XAI)** is rapidly increasing. Many modern AI models, especially deep learning and ensemble methods, operate as **black boxes** where inputs and outputs are visible but the internal reasoning is hidden. This lack of interpretability creates hesitation among users, regulators, and developers. XAI addresses this gap by making AI decisions understandable, transparent, and trustworthy.
The need for XAI arises from several practical, ethical, and legal reasons described below.

Trust

Trust is the primary requirement for adopting any intelligent system. When users do not understand *why* a model produced a particular result, they are less likely to rely on it, even if the result is accurate.

For example, if an AI system predicts that a patient is at risk of heart disease but does not explain the contributing factors (such as blood pressure, cholesterol, age, etc.), doctors may hesitate to act on the recommendation. Similarly, a bank customer denied a loan by an AI system would want to know the reason behind the rejection.

Explainable AI builds **confidence** by clearly showing the factors influencing decisions, allowing users to verify and accept outcomes.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Transparency

Transparency refers to the visibility of how an AI model processes input data and reaches conclusions. Complex models like neural networks involve millions of parameters, making it difficult to trace their behavior.

Without transparency:

- Hidden errors may remain undetected.
- Incorrect correlations may be learned by the model.
- Overfitting or data leakage problems may go unnoticed.

XAI provides tools such as feature importance, decision paths, attention maps, and rule extraction to make model behavior observable. This helps stakeholders understand *what the model is actually learning* rather than assuming it works correctly.

Bias Detection and Fairness

AI models learn from historical data. If this data contains biases related to gender, race, age, or socio-economic status, the model can unknowingly propagate these biases.

For instance:

- Hiring algorithms may prefer male candidates if trained on biased data.
- Credit scoring systems may unfairly reject applications from certain communities.
- Facial recognition systems may perform poorly on certain ethnic groups.

Explainable AI allows developers and auditors to inspect **which features influence decisions**, helping to detect unfair patterns and take corrective action. This is essential for building **fair and ethical AI systems**.

Regulatory and Legal Requirements

Governments and regulatory bodies worldwide are introducing laws that require transparency in automated decision-making systems.

A well-known example is the **General Data Protection Regulation (GDPR)** in the European Union, which includes the "**right to explanation**." This means individuals have the right to know how automated systems made decisions that affect them.

Other sectors such as healthcare, finance, and autonomous systems also require explainability for compliance and certification. XAI helps organizations meet these legal requirements by

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

providing interpretable outputs and documented decision logic.

Debugging and Model Improvement

For AI developers and researchers, understanding why a model makes certain predictions is crucial for improving performance.

Without explainability:

- It is difficult to identify why the model fails in certain cases.
- Incorrect feature learning cannot be easily detected.
- Model tuning becomes a trial-and-error process.

XAI tools help developers analyze errors, refine features, remove irrelevant inputs, and retrain models effectively. This leads to **better accuracy and robustness**.

Human–AI Collaboration

In many applications, AI is not meant to replace humans but to assist them. For effective collaboration, humans must understand AI reasoning.

For example:

- Doctors working with diagnostic AI
- Judges reviewing AI-based risk assessments
- Engineers monitoring AI-controlled IoT systems

Explanations allow humans to validate, override, or refine AI decisions, creating a cooperative environment rather than blind automation.

Risk Management and Safety

In safety-critical systems like autonomous vehicles, industrial IoT, and robotics, wrong decisions can lead to severe consequences. Understanding how AI reaches decisions helps in risk assessment and system validation.

Explainability enables system designers to verify that the AI behaves logically under different scenarios and does not rely on spurious correlations.

INTERPRETABILITY VS EXPLAINABILITY

In Explainable AI literature, the terms **interpretability** and **explainability** are often used as if they mean the same thing. However, there is a subtle but important difference between them.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Understanding this difference is necessary when designing AI systems that are both transparent and user-friendly.

Both concepts aim to make AI decisions understandable, but they differ in **how** and **at what stage** this understanding is achieved.

Interpretability

Interpretability refers to the **degree to which a human can directly understand the internal mechanics of a model without any additional tools**. In interpretable models, the relationship between input features and output predictions is clear by design.

These models are usually simple, structured, and mathematically transparent.

Examples of interpretable models:

- Linear Regression
- Logistic Regression
- Decision Trees
- Rule-based systems
- k-Nearest Neighbors (to some extent)

For instance, in a decision tree used for loan approval, a user can easily trace the path:
If income > ₹50,000 and credit score > 700 → Approve loan

Here, the logic is directly visible. No extra explanation method is needed because the model itself is understandable.

Key point:

Interpretability is **built into the model structure**.

Explainability

Explainability refers to the **ability to explain the decisions of complex or black-box models using additional methods or tools**. These models are not naturally understandable, so external techniques are used to generate explanations.

Examples of black-box models that require explainability:

- Deep Neural Networks
- Random Forest

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

- Gradient Boosting
- Support Vector Machines

For example, a deep learning model predicting a disease from medical images may have millions of parameters. It is impossible to directly understand its internal working. Techniques like:

- LIME (Local Interpretable Model-Agnostic Explanations)
- SHAP (SHapley Additive Explanations)
- Saliency maps
- Attention visualization

are used to explain which parts of the image influenced the decision.

Key point:

Explainability is **added on top of the model**.

Aspect	Interpretability	Explainability
Meaning	Ability to understand internal model directly	Ability to explain output after model decision
Model Type	Simple models (linear, tree)	Complex models (deep learning, ensembles)
Approach	Intrinsic	Post-hoc
Example	Decision tree rules	LIME, SHAP explanations

TYPES OF MODEL INTERPRETABILITY

Model interpretability in Explainable AI can be broadly categorized into two major types based on **when** and **how** understanding of the model is achieved:

1. **Intrinsic (Built-in) Interpretability**
2. **Post-hoc Explainability**

These two approaches address the same goal—making AI understandable—but they differ in methodology and applicability.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

4.1 Intrinsic Interpretability

Intrinsic interpretability refers to models that are **transparent by design**. Their structure, mathematical formulation, and decision process are simple enough for humans to understand without any additional explanation tools.

These models clearly show how input features contribute to the output. A user can trace the decision logic directly from the model itself.

Examples of intrinsically interpretable models:

- **Linear Regression** – Shows how each feature linearly contributes to the output through coefficients.
- **Logistic Regression** – Indicates how input variables affect the probability of a class.
- **Decision Trees** – Provide a tree-like structure where each decision path is visible.
- **Rule-Based Models** – Use human-readable IF–THEN rules.

Characteristics:

- Easy to visualize and understand
- Transparent mathematical relationships
- Suitable for applications requiring high accountability
- No need for extra explanation techniques

Example:

In a decision tree predicting disease risk:

If age > 60 and BP > 140 → High risk

The reasoning is directly readable from the model.

Limitation:

Although these models are easy to understand, they may not perform well on **highly complex, high-dimensional data** such as images, speech signals, or large IoT sensor streams.

4.2 Post-hoc Explainability

Post-hoc explainability refers to explanation methods that are applied **after a complex model has been trained**. These methods do not simplify the model itself but provide explanations for its predictions.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

This approach is necessary for **black-box models** such as deep neural networks, ensemble methods, and support vector machines, where direct interpretation is not possible.

Common post-hoc explainability techniques:

- **LIME (Local Interpretable Model-Agnostic Explanations)**
  Explains individual predictions by approximating the complex model locally with a simple interpretable model.
- **SHAP (SHapley Additive exPlanations)**
  Uses game theory to calculate the contribution of each feature to a prediction.
- **Feature Importance Plots**
  Show which features most influence model decisions overall.
- **Saliency Maps**
  Highlight important regions in images that influence predictions.

Characteristics:

- Work with any black-box model (model-agnostic in many cases)
- Provide local or global explanations
- Help users understand predictions without modifying the model
- Useful when high accuracy is required from complex models

Example:

A convolutional neural network diagnosing pneumonia from X-ray images can be explained using a saliency map that highlights the affected lung region.

Limitation:

Post-hoc explanations may **approximate** the model behavior and sometimes do not perfectly represent the true internal logic.

POPULAR XAI TECHNIQUES

5.1 Feature Importance

This shows which features contribute more to prediction.

5.2 LIME

LIME approximates complex model locally with simple model to explain individual prediction.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

5.3 SHAP

SHAP is based on game theory and provides contribution of each feature.

5.4 Saliency Maps

Used in image models to highlight important pixels.

5.5 Rule Extraction

Rules are extracted from trained neural networks.

Table 1: Comparison of XAI Techniques

Technique	Model Agnostic	Local/Global	Suitable For	Complexity
LIME	Yes	Local	Text, Image, Tabular	Medium
SHAP	Yes	Both	All data types	High
Feature Importance	No	Global	Tabular	Low
Saliency Maps	No	Local	Image	Medium
Rule Extraction	No	Global	Neural Networks	High

VISUALIZATION IN XAI

Visualization tools help in understanding complex model behavior.

- Heatmaps
- Partial Dependence Plots (PDP)
- Individual Conditional Expectation (ICE)
- Decision boundary plots

These methods provide graphical explanation.

APPLICATIONS OF XAI

- **Healthcare**

  Doctors need explanation of disease prediction models.

- **Finance**

  Loan approval and fraud detection requires transparency.

- **Autonomous Vehicles**

  Understanding decision of self-driving cars is very important.

- **Education**

  AI based grading systems must justify evaluation.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

CHALLENGES IN EXPLAINABLE AI

- Trade-off between accuracy and interpretability
- Computational overhead of explanation
- Lack of standard metrics
- User understanding of explanations
- Scalability for big models

Evaluation Metrics for XAI

Some metrics are used to evaluate explanations:

- Fidelity
- Consistency
- Stability
- Human evaluation
- Completeness

REGULATORY AND ETHICAL ASPECTS

Governments are focusing on ethical AI. GDPR mentions “right to explanation”. Organizations must ensure AI fairness and accountability.

FUTURE DIRECTIONS

- Explainability for large language models
- Real time explanation systems
- Human-AI collaborative explanations
- Standard benchmarks for XAI

CONCLUSION

Explainable AI is becoming a critical requirement for modern intelligent systems. As AI models become more complex, need for transparency and trust also increases. Various XAI methods like LIME, SHAP, visualization, and rule extraction help to understand black box models. Despite many challenges, research in this field is progressing rapidly. In future, XAI will be integrated as mandatory component of AI systems rather than optional feature.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved

REFERENCES

1. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier.
3. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions.
4. Molnar, C. (2020). Interpretable Machine Learning.
5. Lipton, Z. C. (2018). The mythos of model interpretability.
6. Guidotti, R., et al. (2018). A survey of methods for explaining black box models.
7. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence.
8. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on XAI.
9. Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts and challenges.
10. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey.

Page 58-68 © MANTECH PUBLICATIONS 2025. All Rights Reserved