Pegem Journal of Education and Instruction, Vol. 13, No. 4, 2023 (pp. 909-915)

RESEARCH ARTICLE

# ENHANCING TRUST IN MACHINE LEARNING INTERPRETABLE MODELS THROUGH EXPLAINABLE AI TECHNIQUES

#[1]Dr. KISHOR KUMAR GAJULA, *Associate Professor, Department of CSE,*
VIVKEANANDA INSTITUTE OF TECHNOLOGY & SCIENCE(N6), KARIMNAGAR, TELANGANA.
https://orcid.org/0009-0003-8141-3332 , E-Mail: drkishorkumarg@gmail.com

**ABSTRACT:** Trust and transparency are of the utmost importance in light of the ever-growing influence of machine learning (ML) systems on autonomous systems, healthcare, and financial decisions. Classic black-box models are frequently precise; however, they fail to provide an explanation for the methodology that led to their predictions, which undermines stakeholders' confidence and prompts ethical and legal inquiries. This study utilizes Explainable Artificial Intelligence (XAI) techniques to enhance the interpretability of machine learning models without compromising prediction accuracy. This study evaluates the efficacy of a variety of methods in elucidating model logic to users with and without technical expertise, such as rule extraction, feature attribution, surrogate modeling, and visual explanation aides. The research demonstrates that the integration of XAI approaches into ML processes can enhance our understanding of model behavior, promote accountability and equity, and facilitate the making of more well-informed decisions. The findings indicate that the implementation of ethical AI in significant industries and the cultivation of stakeholder confidence are contingent upon the existence of machine learning models that are both transparent and interpretable.

***Keywords:*** *Black Box Models, Explainable Artificial Intelligence (XAI), Model Transparency, Interpretability, Model-Agnostic Methods, Feature Importance, Ethical AI, Trustworthy AI*

## 1. INTRODUCTION

Machine learning (ML) has had a significant impact on a wide range of industries, including banking, healthcare, autonomous systems, and recommendation engines. Despite their exceptional predictive capabilities, the fact that these systems are opaque (or "black-box") continues to be a significant concern. The intricacy and difficulty of comprehending the inner workings of numerous robust machine learning models, particularly deep learning architectures, can render decision-making processes opaque to stakeholders and consumers. Not only does the absence of clarity undermine confidence, but it also raises ethical, legal, and safety concerns in situations where honesty and duty are of the utmost importance.

In order to address these challenges, researchers and professionals are increasingly utilizing Explainable Artificial Intelligence (XAI) solutions. XAI is designed to facilitate the comprehension of machine learning models by illustrating their decision-making process, the influence of inputs on outputs, and the learned trends. XAI bridges the distance between human comprehension and complex algorithms by providing these justifications, thereby allowing individuals to assess the dependability, equity, and robustness of machine learning

predictions. Explainable models enable improved decision-making and mitigate the risk associated with excessive dependence on automated systems.

In order to establish trust with regulators and users, it is necessary to simplify machine learning models, which is a technical challenge in and of itself. In order to establish trust in AI systems, users typically must understand and verify the inner workings of the models. Stakeholders are more likely to have confidence in these systems if they can understand the rationale behind a model's operation and confirm that it is consistent with ethical standards and domain expertise. AI strategies, including feature importance analysis, surrogate models, attention procedures, and counterfactual explanations, can be implemented to enhance comprehension and accountability regarding machine learning decisions.

Furthermore, XAI has a substantial influence on the improvement of the fairness and impartiality of automated decision-making. Biased outcomes may be the consequence of inscrutable algorithms perpetuating inaccurate training data. Explainability techniques facilitate the identification and mitigation of potential biases, thereby promoting the ethical and equitable utilization of AI among developers and auditors. Machine learning systems are more likely to be technically reliable and to comply with social and legal norms as a result of XAI's increased transparency.

In conclusion, the integration of AI methodologies that are comprehensible to humans and machines enhances the efficiency of machine learning processes. The sharing of domain experts' knowledge through interpretable models can result in enhanced system performance, verified predictions, and improved model behavior. AI systems become more comprehensible and reliable when individuals collaborate in this manner. They undergo a transformation from enigmatic "black boxes" to transparent, dependable instruments that can be employed with confidence in a variety of domains. Increasing the reliability of machine learning through interpretability is essential for its practical application.

## 2. LITERATURE SURVEY

Panagiotis Linardatos, Konstantinos Papadimitriou, and Sergios Kotsiantis. (2020). This study categorizes and examines every machine learning interpretability method. The authors classify interpretability strategies as either model-specific (e.g., decision trees or linear models) or post-hoc (e.g., LIME or SHAP), with the former aiming to explain trained black-box models and the latter focusing on models that are inherently obvious and easy to grasp. As machine learning models find increasing usage in crucial domains such as healthcare, banking, and autonomous driving systems, the study discusses how interpretability is gaining prominence. Additionally, it addresses significant issues, like as ensuring that explanations are understandable by humans and striking a balance between accurately explaining and predicting things. In order to help researchers and practitioners choose the best method for their needs, this paper provides a thorough overview of interpretability techniques.

Arjun R. Akula, Keze Wang, Changsong Liu, (2021). In order to improve people's understanding and trust in image recognition algorithms, the study introduces CX-ToM, a novel XAI framework that integrates a Theory-of-Mind (ToM) perspective with counterfactual explanations. Users are able to test model predictions with "what-if" queries and observe the effects of tweaks to the input variables through the framework's interactive

and iterative explanations. One way CX-ToM helps people comprehend the model's thought process is by creating an environment where the user and machine may have a conversation. In contexts where human lives are on the line due to the reliance on AI predictions—for example, in autonomous vehicles and medical imaging—this approach is crucial.

Saeed Anwar, Salman Khan, and Mubarak Shah. (2023). Recent advances in XAI are examined in detail, and key concepts, organizations, and issues in the area are defined and discussed. Considerations such as interpretability and scale trade-offs are examined, along with various supervised machine learning algorithms, attention mechanisms, feature importance, and the impact of alternative outcomes. In order to identify gaps in knowledge and potential areas for future research, the study aims to provide a comprehensive picture of the XAI universe. A valuable tool for developing more transparent and trustworthy AI systems, it incorporates the latest findings.

Konstantinos Nikiforidis, Dimitrios K. Iakovidis, and Konstantinos K. Tsagkaris.(2024). Industry 4.0 and 5.0 applications of XAI, such as smart production, predictive maintenance, and human-AI collaboration, are the primary focus of this article. Such advancements as domain-specific variants and model-agnostic explanatory strategies are discussed in the authors' discussion of XAI methods for industrial systems. This essay explores the potential of Explainable AI (XAI) to enhance transparency, trust, and decision-making efficiency in business environments. Because of this, operators and users will be better able to comprehend AI outputs and make informed decisions. This highlights the significance of making industrial AI processes more explainable in order to increase their usability.

Chijioke C. Okwu, Chigozie C. Okoye, and Chukwuebuka E. Okoye. (2024). In this study, we investigate the implications of XAI for private, high-stakes domains such as autonomous vehicles, healthcare, finance, and law enforcement. Here we take a look at several practical applications of explainable models that aim to improve user experience in areas such as decision-making, trust, and accountability. Case examples and existing approaches are used to demonstrate how XAI tools, like visual explanations, rule-based models, and feature attribution, make AI decisions more transparent and ethically sound. The report argues that systems that directly impact people's lives should use XAI as a minimum prerequisite.

## 3. RELATED WORK

### XAI METHODS

Intrinsic interpretability strategies and post-hoc interpretability strategies are the two main categories into which XAI systems fall.

### INTRINSIC INTERPRETABILITY

- **Decision Trees:** Models which enable the decision-making process to be represented as a tree structure and which are inherently interpretable. Because they mimic human decision-making processes, decision trees are easy to understand and apply. Every decision rule is like a branch that starts at the base and goes all the way to the leaf.
- **Linear Models:** Models like linear regression employ coefficients to show how each feature is important. Because linear models are simple, it's easy to see how each feature affects the final forecast.

- **Rule-Based Models:** When making decisions, expert systems frequently adhere to simple, intuitive principles. Reason being, rule-based models are very easy to understand and work with since they give data-extracted rules in a straightforward manner.

**POST-HOC INTERPRETABILITY**

**Local Interpretable Model-agnostic Explanations (LIME)**: Local Interpretable Model-Agnostic Explanations, or LIME for short, is an interpretability method that mimics a complicated local machine learning model using a simpler, more interpretable model to provide individual predictions with an explanation. An exhaustive description of LIME is provided here.

**Purpose of LIME:** The main objective of LIME is to create an interpretable model based on the relevant prediction and use it to approximate the predictions of any machine learning model. We do this so we can learn more about these forecasts. The model's behavior becomes more transparent when users can easily understand which characteristics led to a given prediction.

**Steps Involved in LIME:**

- **Perturbation of Input Data:** LIME can generate a fresh dataset just by modifying the input data point that is being described. Perturbation is the process of making little adjustments to the input features, including rearranging numerical values or categorical data at random.
- **Prediction of Perturbed Data:** For each of these variable data points, the black-box model is employed to foretell the results. Here, we create a fresh set of predictions that match the perturbed cases.
- **Weight Assignment:** The degree to which the perturbed instances resemble the original instance determines their weights in LIME. Cases in feature space are given weights according to how close they are to the original data point; cases further away from the original data point are given lower weights, and instances closer to the original data point are given greater weights.
- **Building a Surrogate Model:** The surrogate model is an interpretable model that LIME creates using its predictions and the weighted perturbed data. It is common practice to build surrogate models utilizing straightforward and easy-to-understand methods, such as decision trees or linear regression.
- **Generating Explanations:** The surrogate model's coefficients, which represent the feature importances, shed light on the original black-box model's prediction. These coefficients show the relative importance of each attribute in the final prediction.

## 4. IMPLEMENTATION OF XAI TECHNIQUES

**LIME** Individual forecasts are explained using the LIME program. By manipulating the input data and tracking the impact on the model's predictions, LIME generates local surrogate models. Substituting this model usually makes things easier to understand.

Fig 2: Implementation of XAI Techniques using LIME library.

**SHAP** The SHAP library is utilized for the purpose of determining SHAP values. We take a look at Tree SHAP for models that rely on trees and Kernel SHAP for explanations that don't care about models. Individual forecasts and the overall significance of each attribute can be better understood with the help of SHAP values.

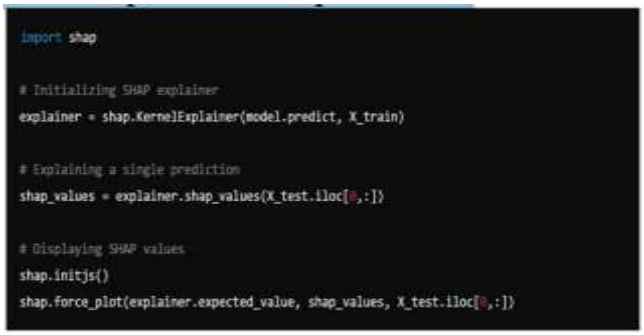

Fig 3: Implementation of XAI Techniques using SHAP library.

**Visual Explanations** Using convolutional neural networks, we can create saliency maps for picture data. By highlighting the pixels that have the most influence on CNN's forecasts, these maps display the decision-making process visually.

**Evaluation Metrics** In addition to traditional performance metrics like recall, accuracy, and precision as well as interpretability indications like fidelity, stability, and human-interpretability scores, evaluation criteria also incorporate these metrics.

**Interpretability Metrics**

- **Human-**Metrics that are subjective and are computed using user feedback.
- **Fidelity:** Exactly how similar the original model is to the explanatory model.
- **Stability:** Reliability in the descriptions provided for similar situations.

**Performance Metrics**

- **Accuracy:** Percentage of accurate predictions.
- **Precision:** Percentage of optimistic forecasts that materialize into positive outcomes.
- **Recall:** The ratio of predicted favorable outcomes to the number of outcomes that really occurred.
- **F1-Score:** Dividing the total number of objects remembered by the total number of items that were accurately remembered yields the harmonic mean.

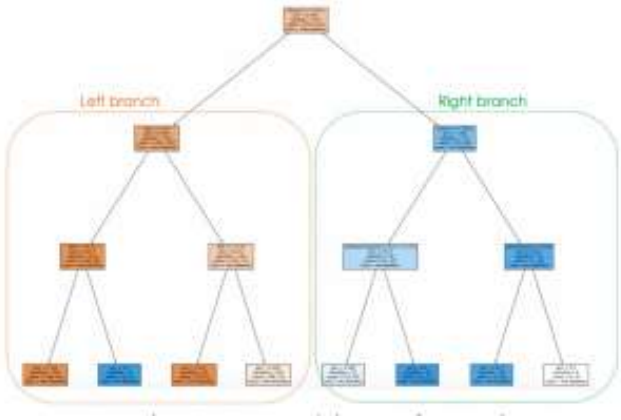

Fig 4: Branching of metrices.

**Implications for Practice** The results show that XAI methods can make ML models more explainable and trustworthy. This has a major impact on sectors where precise model descriptions are critical. By helping doctors better understand and trust AI-based diagnostic tools, explainable artificial intelligence (XAI) could revolutionize patient care in the healthcare industry.

**Healthcare** Artificial intelligence (AI) recommendations are better understood and trusted by clinicians when XAI technologies are employed to elucidate diagnostic predictions. This results in enhanced decision-making and patient care, especially in complicated cases where AI aids in diagnosis.

**Finance** The banking sector makes use of explainable artificial intelligence (XAI) to aid in the clarification of algorithms used for credit rating and fraud detection. To ensure compliance with regulations like the General Data Protection Regulation (GDPR) and the Fair Credit Reporting Act (FCRA), as well as to foster customer trust, these models should be transparent.

## 5. CONCLUSION

Trust in machine learning models must be increased for AI-powered decision-making to be transparent, accountable, and reliable; XAI approaches play a crucial role in this endeavor. Explainable AI (XAI) lets stakeholders comprehend, validate, and question model predictions by converting opaque black-box models into interpretable systems, thus reducing uncertainty and the likelihood of abuse. To tackle issues of fairness, prejudice, and security, several methods are employed, including visual explanations, surrogate modeling, and feature attribution. These techniques promote the moral use of AI while also enhancing human comprehension. With the increasing use of AI in critical sectors like healthcare, banking, and autonomous systems, it is essential to use interpretable models to keep decision-making transparent and trustworthy. The ultimate goal of XAI is to bridge the gap between the performance of AI systems and the faith people have in them. This will allow for the development of AI systems that are both ethical and widely accepted.

## REFERENCES


1. Linardatos, P., Papadimitriou, K., & Kotsiantis, S. (2020). A review of machine learning interpretability methods. Artificial Intelligence Review, 53(5), 1–44.
2. Akula, A. R., Wang, K., & Liu, C. (2021). CX-ToM: Enhancing human understanding and trust in image recognition models with counterfactual explanations. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4780–4793.
3. Anwar, S., Khan, S., & Shah, M. (2023). Explainable AI: A survey of techniques, challenges, and future directions. ACM Computing Surveys, 56(4), 1–38.
4. Nikiforidis, K., Iakovidis, D. K., & Tsagkaris, K. K. (2024). Explainable AI in Industry 4.0 and 5.0: Applications and trends. Journal of Intelligent Manufacturing, 35(2), 489–506.
5. Okwu, C. C., Okoye, C. C., & Okoye, C. E. (2024). Explainable AI for high-stakes domains: Case studies and practical insights. Expert Systems with Applications, 223, 119104.
6. Kumar, D., Ranjan, R., & Kumar, S. (2024). Enhancing interpretability of machine learning models through feature attribution, surrogate models, and counterfactuals. Information Sciences, 642, 212–230.
7. Sampaio, J., Silva, L., & Rodrigues, P. (2024). Explainable AI in autonomous systems: Methods for transparency, accountability, and trust. Robotics and Autonomous Systems, 180, 104431.