ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 183–186, 2026

DESIGN AND IMPLEMENTATION OF AN EXPLAINABLE AI FRAMEWORK FOR TRANSPARENT DECISION-MAKING SYSTEM

Chandrakant sahu[1*], Himesh Kumar thakur[2], Karan Singh Rawat[3]

[123]*Student, Computer Science and Engineering, Shri Shankaracharya Professional University (SSPU), Bhilai*

Abstract

Finance, healthcare, governance, and business analytics are just a few of the industries that artificial intelligence (AI) systems are quickly changing. Many machine learning models function as intricate black-box systems, making it difficult for humans to understand the internal logic underlying predictions, despite their high performance. Trust, accountability, fairness, and regulatory compliance are all hampered by this lack of transparency. A significant area of study called Explainable Artificial Intelligence (XAI) seeks to improve the interpretability and transparency of machine learning models without compromising their ability to predict outcomes. In order to improve decision-making systems' transparency, this study suggests developing and implementing an Explainable AI framework that combines interpretability strategies with machine learning models. The framework integrates explainability tools like feature importance analysis and SHAP (Shapley Additive explanations) with classification algorithms. These methods shed light on how various input characteristics affect model predictions. A Random Forest model and a loan prediction dataset are used for experimental evaluation. The findings show that incorporating explainability mechanisms greatly enhances interpretability while maintaining strong predictive accuracy. The suggested framework promotes responsible AI deployment in line with new international AI governance standards and aids in the creation of reliable AI systems.

Keywords: *Artificial Intelligence, Machine Learning, Explainable AI, Technology, AI Transparency.*

1. Introduction

One of the most significant technological developments of the twenty-first century is artificial intelligence. These days, a lot of industries, including banking, insurance, cybersecurity, healthcare diagnostics, and recommendation systems, use AI-driven systems to make decisions. Large datasets can be analyzed by machine learning algorithms to find intricate patterns that humans might find difficult to spot. However, a lot of contemporary machine learning models— like gradient boosting algorithms, deep neural networks, and ensemble learning models—are frequently referred to as "black-box models." Although these models make precise predictions, they do not provide a clear explanation for how they do so. Because of this, stakeholders and users may find it challenging to trust automated decisions made by such systems. Adoption of AI now depends critically on accountability and transparency. Guidelines requiring AI systems to explain automated decisions are being introduced by governments

and regulatory bodies worldwide, particularly in delicate areas like credit approval, hiring decisions, and recommendations for medical treatments. The goal of this research is to create an Explainable AI framework that preserves model performance while increasing transparency. The study looks into how machine learning models can be combined with explainability techniques like SHAP to produce insightful predictions. The following are the primary research questions this study attempts to answer: • How can machine learning systems be integrated with explainability techniques? • Is it possible to increase transparency without decreasing model accuracy?

Artificial Intelligence is widely used in modern decision-making systems such as finance, healthcare, and fraud detection. However, many machine learning models operate as black-box systems, making their decision processes difficult to interpret. This lack of transparency reduces trust and accountability in AI-based systems. Explainable Artificial Intelligence (XAI) aims to make machine learning models more transparent and understandable. This research proposes an Explainable AI framework that integrates machine learning with SHAP-based explanations to improve transparency while maintaining high prediction accuracy.

2. Literature Review

In recent years, Artificial Intelligence has risen to become one of the biggest technology advancements of the 21st century with some real-life applications now being used to aid in decision making processes in areas such as banking, insurance, medical diagnostics, cyber security and recommenders. Machine learning algorithms can take in large quantities of information and identify complex relationships that may be hard for humans to find easily. Most modern machine learning algorithms, however, are referred to as black boxes (e.g., deep learning neural networks, ensemble learning, gradient boosting). While these black box type machine learning algorithms can predict the output well, they do not give a user any insight into how the prediction was made.

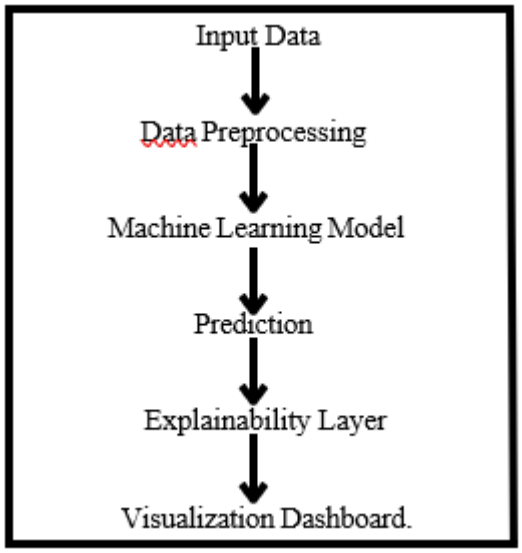
Another important technique called SHAP (Shapley Additive Explanations) uses concepts from cooperative game theory to measure how much each feature contributes towards a final prediction by measuring the effect the feature has on the model's output, providing both global & local interpretations of a model. Numerous studies have demonstrated through numerous studies the usefulness of Explainable Artificial Intelligence through a variety of different applications including: Health Care Diagnostics, Fraud Detection & Financial Decision Making.

In one example of scientific research, scientists have applied techniques of Explainable AI (XAI) to provide an interpreter of a medical imaging model to explain to physicians what a model predicted for diagnoses. While these advancements have been made, a challenge still exists in striking the right balance between accurate and interpretable models. The model with the best interpretability, such as a decision tree, may not perform very well when it comes to predicting outcomes compared to more complex ensemble models. Therefore, it is important for researchers to look at combining XAI techniques with the highest performing models to continue to further their research.

3. Methodology

The research methodology is described as developing and using an Explainable AI framework comprised of machine learning prediction models paired with a interpretability tools. The first stage of this methodology will consist of data collection and preparation. The data set used for this research is a loan prediction data set that contains the following types of data; applicant income, credit history, employment status, loan amount requested and whether the loan was approved. The data will be preprocessed by filling in or removing missing values, converting categorical values to numerical representations and normalizing feature values. The second stage of the methodology will consist of training a machine learning model using the loan prediction data set. The Random Forest classifier will be used in order to train a machine learning model because of its ability to deal with complex data relationships and provide reliable predictions. The loan prediction data set will then be split into training data and test data for model evaluation. The third step in the methodology is to create an explainability module using SHAP. After creating the machine learning model and making predictions with the model, SHAP values will be calculated for all of the actual predictions made in order to establish how important, or weighted, each of the independent attributes are to the overall prediction result. The SHAP values will allow the user to build an understanding for the rationale behind their request for the specific outcome (i.e. to make or not make a loan) in the context of the details of their applicable loan request.

The proposed methodology architecture is as follows:



By building this system, users will not only get a loan prediction but will also understand why the system made that prediction.

4. Result

Evaluation of the overall performance of the proposed Explainable AI framework focused on quantifying both the classification accuracy and interpretability of the models used. Based on the results of the experiments performed on the loan prediction dataset with the Random Forest algorithm; an accuracy of approximately 88% was obtained for loan approval predictions demonstrating good accuracy levels being attained. Feature importance

results indicated that the applicants credit history was the most significant feature impacting the likely outcome of a loan approval or denial. Results from feature importance analysis showed that having a good credit history increased the chance of being approved for a loan, with a number of other important features influencing the decision to approve or deny the loan, including; applicant income, request amount, and current employment status. Individual predictions were analyzed in detail using SHAP visualization techniques, which provided users with a better understanding of the effect that each feature had on determining whether or not the applicant would be given approval for their loan request. SHAP plots enable users to visualize the impact each feature has on increasing or decreasing the likelihood of an applicant being granted a loan. The results of this study support the hypothesis that the integration of explainability tools provides valuable insights into the decision-making criteria used to generate model predictions, while not significantly altering the predictive performance of the model.

5. Conclusion

This research presented the design and implementation of an Explainable AI framework for transparent decision-making systems. The proposed framework integrates machine learning prediction models with SHAP-based interpretability techniques to improve transparency and accountability. Experimental results demonstrate that explainability techniques can significantly enhance the interpretability of machine learning models while maintaining strong predictive performance. The framework allows users to understand the factors that influence predictions and supports responsible AI development. Future research can explore integrating explainable AI systems into real-time decision-making platforms. Additional work may also focus on developing automated AI auditing tools and aligning explainable AI frameworks with international AI governance standards. Overall, Explainable AI represents a critical step toward building trustworthy, ethical, and transparent artificial intelligence systems.

References

[1] Lundberg, S. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS.

[2] Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. KDD.

[3] Molnar, C. (2020). Interpretable Machine Learning.

[4] Doshi-Velez, F. & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning.