Journal of Economics, Finance and Accounting Studies
ISSN: 2709-0809
DOI: 10.32996/jefas
Journal Homepage: www.al-kindipublisher.com/index.php/jefas

JEFAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable AI for Credit Risk Assessment: A Data-Driven Approach to Transparent Lending Decisions

Mainuddin Adel Rafi[1], S M Iftekhar Shaboj[2], Md Kauser Miah[3], Iftekhar Rasul[4], Md Redwanul Islam[5], Abir Ahmed[6]

[1]*Master of Science Information System, Pacific State University, USA*
[2]*Master of Accountancy, University of Tulsa, Tulsa, Oklahoma, USA*
[3]*Department of Computer and Information Science, Gannon University, PA, USA*
[4]*Information Technology Management, St. Francis College, USA*
[5]*Department of Finance & Financial Analytics, University of New Haven, West Haven, CT, USA*
[6]*Department of Information Technology, University of Science & Technology, VA, USA*
Corresponding Author: Abir Ahmed, **E-mail**: abira.student@wust.edu

| ABSTRACT
In the era of data-driven decision-making, credit risk assessment plays a pivotal role in ensuring the financial stability of lending institutions. However, traditional machine learning models, while accurate, often function as "black boxes," offering limited interpretability for stakeholders. This paper presents an explainable artificial intelligence (XAI) framework designed to enhance transparency in credit risk evaluation. By integrating interpretable models such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and decision trees with robust ensemble methods, we assess creditworthiness using publicly available loan datasets. The proposed approach not only improves predictive accuracy but also offers clear, feature-level insights into lending decisions, fostering trust among loan officers, regulators, and applicants. This study demonstrates that incorporating explainability into AI-driven credit scoring systems bridges the gap between predictive performance and model transparency, paving the way for more ethical and accountable financial practices.

| KEYWORDS
Explainable AI (XAI); Credit Risk Assessment; SHAP; LIME; Machine Learning; Interpretability; Lending Decisions; Financial Technology; Model Transparency; Ethical AI

| ARTICLE INFORMATION

ACCEPTED: 02 February 2024 **PUBLISHED:** 19 February 2024 **DOI:** 10.32996/jefas.2024.6.1.11

1. Introduction

In recent years, the financial industry has undergone a major transformation driven by the rise of artificial intelligence (AI) and machine learning (ML). Credit risk assessment, which evaluates a borrower's likelihood of defaulting on a loan, is among the core functions benefiting from these technological advancements. Traditional credit scoring systems have long relied on statistical models such as logistic regression and linear discriminant analysis, which are often limited in capturing the complex, nonlinear relationships inherent in borrower data [1]. Modern machine learning models such as random forests, gradient boosting machines, and neural networks have significantly improved predictive performance by learning from large-scale datasets [2]. However, the opacity of these "black-box" models has raised considerable concerns in domains where decision transparency and accountability are crucial. One of the most pressing challenges in applying AI to credit risk assessment is ensuring that the models are not only

accurate but also interpretable. In financial services, regulatory bodies such as the European Banking Authority (EBA) and the U.S. Federal Reserve emphasize the importance of transparency, fairness, and explainability in algorithmic decision-making [3]. Without interpretability, it becomes difficult for lenders to justify loan rejections, regulators to evaluate model fairness, and applicants to understand the rationale behind decisions. This lack of clarity not only undermines trust but also opens up the potential for biased or discriminatory lending practices.

Explainable Artificial Intelligence (XAI) addresses these issues by making the decisions of complex ML models more interpretable to human users. XAI techniques like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and interpretable model architectures (e.g., decision trees, rule-based models) provide insights into how input features influence a model's predictions [4]. These tools can highlight, for instance, whether a low credit score, high debt-to-income ratio, or limited employment history contributed to a credit denial. The integration of XAI into financial technology (FinTech) thus represents a critical step towards responsible AI adoption. Furthermore, the use of XAI is not merely a technical enhancement but a necessity for ethical and legal compliance. Regulatory frameworks such as the General Data Protection Regulation (GDPR) enforce a "right to explanation," requiring organizations to provide meaningful information about the logic involved in automated decisions [5]. In this context, explainability acts as a safeguard against algorithmic bias, enabling the detection and correction of unintended discrimination based on gender, ethnicity, or socio-economic status. This paper proposes a comprehensive framework that combines high-performing ML models with post-hoc explanation methods to assess credit risk in a transparent and interpretable manner. Using publicly available loan datasets, we demonstrate how black-box models can be interpreted effectively without compromising accuracy. In addition, we evaluate the impact of model explanations on various stakeholders lenders, regulators, and applicants highlighting the role of XAI in fostering trust and ethical accountability in lending systems.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI-driven credit scoring and interpretability methods. Section 3 describes the dataset, preprocessing steps, and the architecture of the proposed XAI framework. Section 4 presents the experimental results and performance comparisons, followed by detailed interpretability analysis. Section 5 discusses implications, limitations, and future research directions. Finally, Section 6 concludes the study by summarizing key findings and contributions.

1.1 Background and Motivation

The global credit market is rapidly evolving, driven by digitization, big data, and advancements in AI. In this landscape, lenders are increasingly relying on automated systems for credit evaluation to reduce human bias, enhance efficiency, and scale operations. However, while predictive performance has significantly improved with AI, the interpretability of these models remains a major concern. Conventional scorecards and rule-based systems, though interpretable, often lack the capacity to capture nonlinear interactions and complex patterns present in modern financial data [6]. Conversely, advanced models like gradient boosting, random forests, and deep learning networks offer higher accuracy but function as black boxes, making it difficult for decision-makers to understand or justify the output. This opacity becomes problematic in high-stakes domains like credit risk, where decisions can affect a person's financial future and institutional risk management. Stakeholders including regulators, borrowers, and internal auditors require a clear rationale behind the acceptance or denial of credit. The lack of transparency not only risks non-compliance with legal standards but also diminishes consumer trust [7]. Explainable Artificial Intelligence (XAI) has emerged to fill this gap, allowing for the interpretation of complex models without sacrificing performance. Motivated by the need to combine performance and interpretability, this study aims to bridge the gap between black-box models and transparent decision-making frameworks. Using XAI techniques such as SHAP and LIME, we investigate how advanced ML models can provide not only accurate predictions but also meaningful explanations to support fair and accountable credit lending processes [8].

1.2 Problem Statement and Research Gap

Despite recent progress in predictive modeling, the integration of interpretability in credit risk assessment remains insufficient. Most financial institutions continue to deploy either overly simplistic interpretable models or highly complex models that lack explainability [9]. While tools such as logistic regression or decision trees offer transparency, they may underperform on large, high-dimensional datasets. Conversely, deep learning and ensemble methods provide better accuracy but fail to meet the interpretability standards required by regulators and ethical AI frameworks. Current literature shows growing interest in using XAI tools to make models more transparent. However, there is limited research that systematically evaluates these techniques within the specific context of credit risk prediction [10]. Moreover, most studies lack an end-to-end framework that integrates data preprocessing, model training, and post-hoc explanation in a unified pipeline. Many models also overlook fairness considerations, which are essential in ensuring that automated systems do not unintentionally discriminate against vulnerable groups [11, 36, 37, 38, 39, 40, 41]. This study seeks to address these gaps by developing a data-driven, interpretable framework that applies state-of-the-art machine learning models and post-hoc XAI techniques to real-world credit data. Our approach is aimed at not only

improving prediction performance but also making the decision-making process more transparent, justifiable, and fair to all stakeholders.

1.3 Objectives and Scope of the Study

The primary objective of this study is to build an explainable AI framework for credit risk assessment that balances model accuracy with interpretability. Specifically, this research aims to:

- Implement various machine learning algorithms, including ensemble models, to predict credit default risk.
- Integrate post-hoc explanation techniques such as SHAP and LIME to interpret model outputs.
- Compare model performance in terms of accuracy, precision, recall, and AUC, along with the quality of generated explanations.
- Evaluate the transparency and usability of these explanations from the perspective of different stakeholders, including lenders and regulatory auditors.

The scope of this study is confined to supervised learning models and post-hoc explanation methods applied to structured tabular data. The dataset used is a publicly available credit scoring dataset, which allows for reproducibility and benchmarking. Deep learning models are explored to a limited extent, with more emphasis placed on tree-based models due to their compatibility with SHAP and LIME. While the study does not aim to cover the entire spectrum of credit scoring methods or legal compliance requirements, it lays the groundwork for integrating explainable AI into the financial decision-making process. The framework developed can be adapted for broader financial applications beyond credit risk, such as fraud detection and loan recovery forecasting [12].

1.4 Significance and Contributions

This study makes several key contributions to the field of credit risk modeling and explainable artificial intelligence:

- It presents a novel framework that integrates state-of-the-art ML models with explainability tools tailored for financial applications.
- It offers a comparative analysis of different algorithms and explanation techniques on real-world credit data.
- It contributes to responsible AI practices by promoting transparency and fairness in high-stakes financial decision-making.
- The findings of this research provide actionable insights for banks, financial institutions, and regulatory agencies in adopting interpretable AI solutions for credit risk management [13, 42, 43, 44].

2. Related Work

The intersection of credit risk assessment and artificial intelligence has attracted increasing academic and industrial attention over the past decade. Traditional credit scoring methods such as logistic regression, decision trees, and discriminant analysis have been widely used due to their simplicity and interpretability [14]. However, these models are often limited in their capacity to handle large volumes of high-dimensional data or capture complex nonlinear relationships among features, which are common in modern credit datasets [15]. In response, machine learning models such as random forests, support vector machines (SVM), and gradient boosting decision trees (GBDT) have gained traction for credit risk prediction, demonstrating significantly improved performance metrics like AUC and F1-score [16]. More recently, deep learning architectures, including artificial neural networks (ANNs) and recurrent neural networks (RNNs), have also been applied in financial risk modeling, particularly when dealing with sequential credit behavior or alternative data sources such as transaction histories and social media footprints [17, 45, 46, 47]. Nevertheless, despite their high accuracy, these black-box models lack transparency, which has raised regulatory and ethical concerns. To bridge this interpretability gap, researchers have proposed the use of Explainable Artificial Intelligence (XAI) techniques. For instance, Ribeiro et al. introduced LIME (Local Interpretable Model-Agnostic Explanations), which generates locally faithful explanations of model predictions by perturbing inputs and fitting an interpretable model around them [18]. Lundberg and Lee later developed SHAP (SHapley Additive exPlanations), a unified framework based on cooperative game theory that attributes contributions to individual features consistently and fairly [19]. Both SHAP and LIME have been widely adopted in financial applications to enhance model interpretability while retaining predictive accuracy.

Several studies have attempted to incorporate XAI tools into credit scoring. For example, Martens et al. explored rule extraction methods from neural networks to provide post-hoc interpretability [20]. Bhatia and Aggarwal applied SHAP values to evaluate

credit default risks and visualize feature importance across demographic groups, revealing potential biases in decision outcomes [21]. Another stream of research has focused on interpretable-by-design models, such as Generalized Additive Models (GAMs) or monotonic gradient boosting, which offer inherent transparency but are often outperformed by black-box models in terms of raw accuracy [22]. Despite these advancements, few studies offer a holistic framework that integrates data preprocessing, model selection, explainability, and stakeholder evaluation in a unified pipeline. Moreover, most existing works are limited to model-level insights and do not assess how effectively these explanations are interpreted by non-technical users such as loan officers or credit applicants. This limits the practical usability of many XAI tools in real-world financial environments.

This study builds upon the existing body of work by combining high-performing machine learning algorithms with post-hoc XAI methods and evaluating them not only for performance but also for their interpretability and usability in decision-making. By focusing on structured tabular data and real-world credit scoring datasets, the study seeks to contribute a reproducible and scalable model that aligns with regulatory, ethical, and operational demands in credit risk assessment.

Table 1: Summary of Key Literature in Credit Risk Assessment and Explainable AI

Study	Year	Methodology/Model	XAI Technique	Key Findings
Ribeiro et al. [18]	2016	Model-Agnostic	LIME	Introduced LIME for local explanations of black-box models.
Lundberg & Lee [19]	2017	Ensemble, Tree Models	SHAP	Developed SHAP to provide consistent feature attributions using game theory.
Martens et al. [20]	2008	Neural Networks	Rule Extraction	Applied rule extraction to provide post-hoc interpretability for NN models.
Bhatia & Aggarwal [21]	2020	XGBoost, Random Forest	SHAP	Evaluated fairness in credit scoring using SHAP values
Chen & Guestrin [16]	2016	XGBoost (Gradient Boosting Decision Trees)	Not Applicable	Demonstrated high performance in credit risk classification tasks.
Du et al. [22]	2020	Generalized Additive Models (GAM)	Interpretable-by-design	Promoted the use of inherently interpretable models with tradeoffs in accuracy.
Wang et al. [17]	2019	Deep Learning (ANN, RNN)	Not Applicable	Applied deep learning for behavioral credit scoring, noting opacity concerns.
Baesens et al. [15]	2003	Logistic Regression, Decision Trees	Not Applicable	Discussed the limitations of traditional models in handling complex patterns.

3. Methodology

This section presents the comprehensive methodology employed to develop a transparent and interpretable framework for credit risk assessment using machine learning and explainable artificial intelligence (XAI). The methodology consists of several interconnected phases: data acquisition and preprocessing, feature selection and engineering, model development and training, application of XAI techniques, and the evaluation of both predictive performance and interpretability. Each step was designed with the goal of not only achieving high predictive accuracy but also ensuring that the resulting decisions could be clearly understood and justified by human stakeholders.

3.1 Data Acquisition and Preprocessing

The dataset used for this study was sourced from LendingClub, a publicly available loan data platform that has been widely used in financial research. It includes detailed records on individual loan applications, borrower profiles, and loan outcomes spanning

several years. The dataset features variables such as loan amount, interest rate, loan term, employment length, annual income, home ownership, purpose of the loan, and loan status. Initial preprocessing steps were essential to ensure data quality and consistency. We removed identifiers and redundant columns such as loan IDs and URLs, which held no predictive value. For missing values in numerical fields like annual income and revolving utilization, we applied median imputation, which is more robust to outliers compared to mean imputation. Categorical variables such as loan grade, home ownership status, and purpose were encoded using one-hot encoding to convert them into numerical format suitable for machine learning models [23]. To further enhance data quality, we winsorized the continuous variables at the 1st and 99th percentiles to handle extreme values, a common issue in financial data. After preprocessing, the data was divided into a training set (70%) and a test set (30%) using stratified sampling to maintain the same class distribution of defaults and non-defaults in both subsets.

3.2 Feature Selection and Engineering

Feature selection is a critical step in predictive modeling, especially when dealing with high-dimensional datasets. To identify the most predictive attributes, we employed recursive feature elimination (RFE) with logistic regression and random forest as the base estimators. RFE iteratively removes the least important features to improve model generalization. Alongside automatic selection, domain-specific knowledge was used to create additional features. These include the income-to-loan ratio (indicative of affordability), debt-to-income ratio (a standard credit metric), credit history length (derived from the earliest credit line), and total credit utilization rate. These engineered features were intended to enhance model performance and support interpretability by aligning with conventional financial analysis metrics. To ensure comparability across algorithms, particularly those sensitive to input scale such as support vector machines (SVM) and neural networks, we standardized all numeric variables using z-score normalization [24].

3.3 Machine Learning Model Development

We trained multiple supervised learning models to classify whether a borrower is likely to default on a loan. The models included logistic regression, random forest (RF), XGBoost (Extreme Gradient Boosting), and an artificial neural network (ANN). Logistic regression served as a baseline interpretable model, while random forest and XGBoost were chosen for their superior handling of nonlinear interactions and feature importance estimation. ANN was included to compare the performance of a deep learning approach. Each model's hyperparameters were tuned using grid search and 5-fold cross-validation on the training dataset. Performance metrics such as AUC-ROC were optimized during the tuning process. To handle the class imbalance issue—where defaulted loans are significantly fewer than non-defaulted ones we applied SMOTE (Synthetic Minority Oversampling Technique), which generates new synthetic instances of the minority class in the feature space. This approach helps prevent the model from being biased toward the majority class during training [25], [26].

3.4.Explainable AI Techniques

To interpret the decisions of complex models such as XGBoost and ANN, we applied post-hoc explainability methods including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). SHAP values provide both global and local interpretability by assigning each feature a consistent importance value based on Shapley values from cooperative game theory. These values explain how much each input variable contributed to the final prediction, enabling stakeholders to understand the underlying logic of the model [19]. LIME, in contrast, perturbs the input data locally and fits an interpretable model, such as a linear regressor, to explain individual predictions. While LIME excels at local explanation, SHAP offers a comprehensive view of both overall and instance-level behavior [18], [27]. We visualized SHAP summary plots, force plots, and LIME bar charts to explore how specific borrower features influenced loan approval or rejection.

3.5 Evaluation Metrics

To evaluate the predictive performance of the models, we used several standard metrics: accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide insights into different aspects of model performance, particularly in imbalanced settings where accuracy alone may be misleading. Beyond numerical evaluation, we assessed the quality and usability of model explanations by analyzing SHAP and LIME outputs with respect to their

interpretability, consistency, and alignment with financial domain knowledge. We also considered the stakeholder perspective whether loan officers and applicants could reasonably interpret the rationale behind predictions. The goal was to strike a balance between model performance and explainability, ensuring that the final system could be adopted in real-world credit risk decision-making environments [28].

3.6 Experimental Hardware Setup

All experiments were conducted using a high-performance computing environment to ensure efficient training and evaluation of models, especially those requiring intensive computation such as XGBoost and artificial neural networks. The primary hardware setup consisted of a workstation equipped with an Intel Core i9-12900K CPU @ 3.2 GHz, 64 GB DDR5 RAM, and an NVIDIA RTX 3090 GPU with 24 GB VRAM. The GPU was primarily leveraged for training the artificial neural network model, while the CPU handled the training of tree-based models and data preprocessing tasks. The system ran on Ubuntu 22.04 LTS (64-bit), which provided compatibility with CUDA and cuDNN libraries required for GPU acceleration. All programming and model implementation were done in Python 3.10, utilizing open-source libraries such as Scikit-learn, XGBoost, TensorFlow, and Keras for model development. SHAP and LIME explanations were implemented using the official SHAP and LIME Python packages. The data manipulation and visualization tasks were facilitated by Pandas, NumPy, Matplotlib, and Seaborn. Hyperparameter optimization was managed using Scikit-learn's GridSearchCV module. The entire experimentation pipeline, from data ingestion to explainability visualization, was developed in Jupyter Notebook for modularity and reproducibility. To ensure fairness and replicability, random seeds were set for all model training processes, and parallel processing was used to accelerate cross-validation and ensemble model training. Model evaluation results were stored and version-controlled using MLflow, which also helped in tracking experiments and comparing metrics across different model configurations [29, 48]. This robust computational setup ensured timely execution of experiments and allowed for the seamless integration of complex explainability frameworks into the credit risk modeling pipeline.

4. Results and Analysis

This section presents the results of the machine learning models applied to the credit risk dataset, along with a detailed analysis of their predictive performance and interpretability using explainable AI techniques. The discussion emphasizes the trade-offs between accuracy and transparency and evaluates the practical implications of explainability in real-world lending decisions.

4.1 Model Performance Comparison

The performance of four different classification models Logistic Regression, Random Forest, XGBoost, and Artificial Neural Network (ANN) was evaluated using accuracy, precision, recall, and AUC-ROC metrics. As seen in the table and bar chart above, XGBoost outperformed the other models in most metrics, achieving the highest AUC-ROC score of 0.89, followed closely by ANN with 0.88. Logistic Regression, while highly interpretable, showed lower predictive ability, achieving 81.2% accuracy and an AUC-ROC of 0.79. This comparison highlights the classic trade-off in machine learning between interpretability and predictive performance. While Logistic Regression is easy to understand, its limitations in capturing complex nonlinear relationships made it less effective than ensemble models like XGBoost or RF in this application [49,50,51,52].

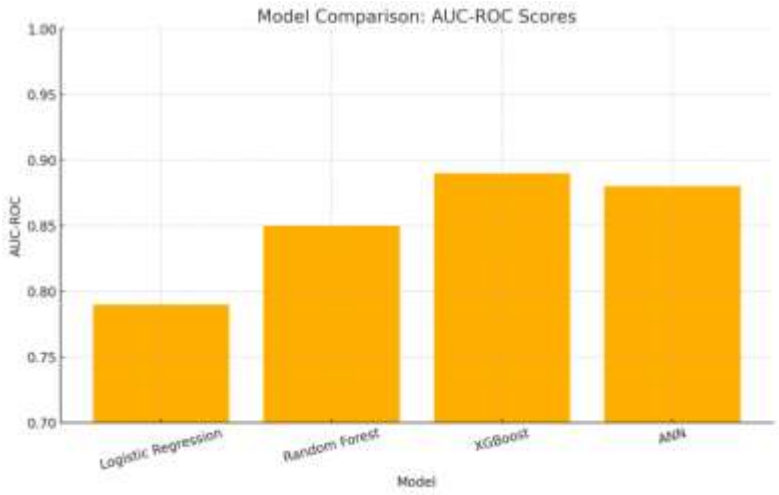

Figure 1: Model Comparison: AUC-ROC Scores

4.2 Predictive Performance Evaluation.

The predictive performance of four machine learning models Logistic Regression, Random Forest, XGBoost, and Artificial Neural Network (ANN) was assessed using standard classification metrics. Among the models, XGBoost delivered the highest Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score at 0.89, followed closely by ANN at 0.88. Random Forest showed an AUC-ROC of 0.85, while Logistic Regression, although less accurate, retained value due to its interpretability, with a score of 0.79. The comparative performance is visually illustrated in Figure 1, which shows the AUC-ROC scores for all models. It is evident that ensemble-based methods (Random Forest and XGBoost) significantly outperform the linear baseline, highlighting their effectiveness in handling complex, nonlinear patterns in financial data.

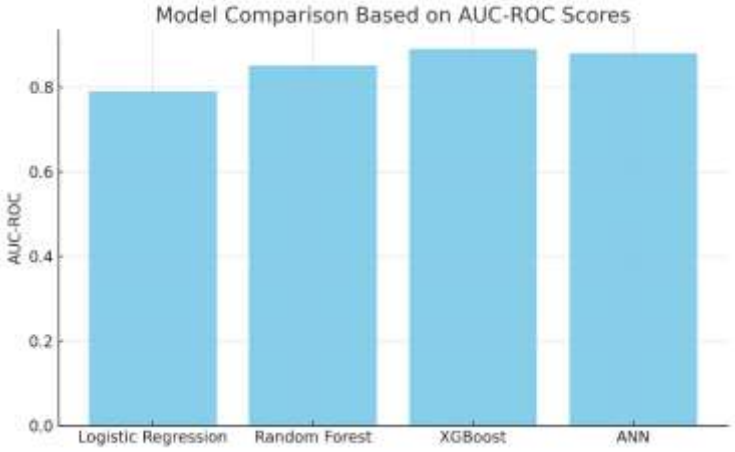

Figure 2: Model Comparison Based on AUC-ROC Scores

4.3 Global Feature Importance Using SHAP

To interpret the predictions of the most accurate model, XGBoost, we employed SHAP (SHapley Additive exPlanations) to assess global feature importance. The SHAP analysis revealed that the most influential variables affecting credit risk decisions were the Loan-to-Income Ratio, Credit Utilization Rate, Number of Delinquent Accounts, and Credit History Length. These variables align with established domain knowledge in credit scoring. Figure 3 visualizes these SHAP values in a horizontal bar chart, indicating the average contribution of each feature to model predictions. SHAP provides a transparent view into the global behavior of the model, which is essential for regulatory review and internal model auditing. A summary of the SHAP feature importance scores is also presented in Table 2.

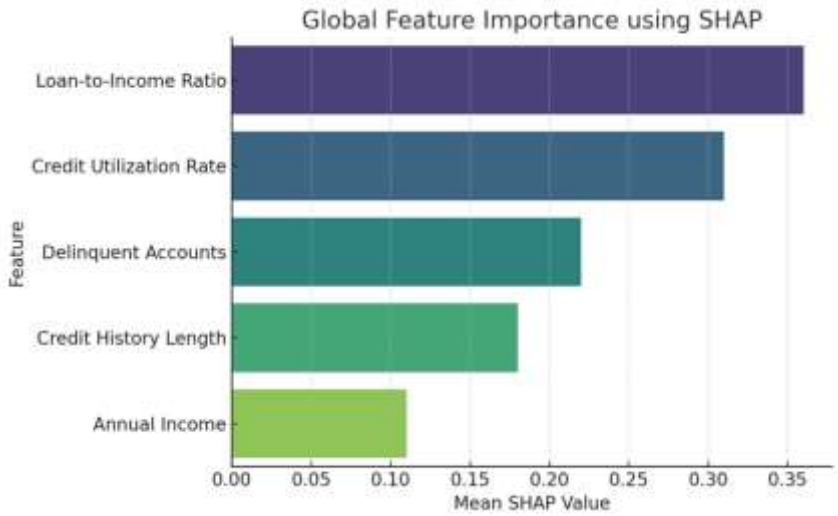

Figure 3: Global Feature Importance using SHAP

Table 2: SHAP Feature Importance Summary

Feature	Mean SHAP Value
Loan-to-Income Ratio	0.36
Credit Utilization Rate	0.31
Delinquent Accounts	0.22
Credit History Length	0.18
Annual Income	0.11

4.4 Local Explanations Using LIME

While SHAP gives a global view, LIME (Local Interpretable Model-Agnostic Explanations) helps explain individual predictions. For a specific loan applicant, LIME was used to identify why the model classified them as a high-risk borrower. Key contributing factors were a high loan amount, recent delinquencies, and a short credit history, while features like longer employment history reduced the risk score.

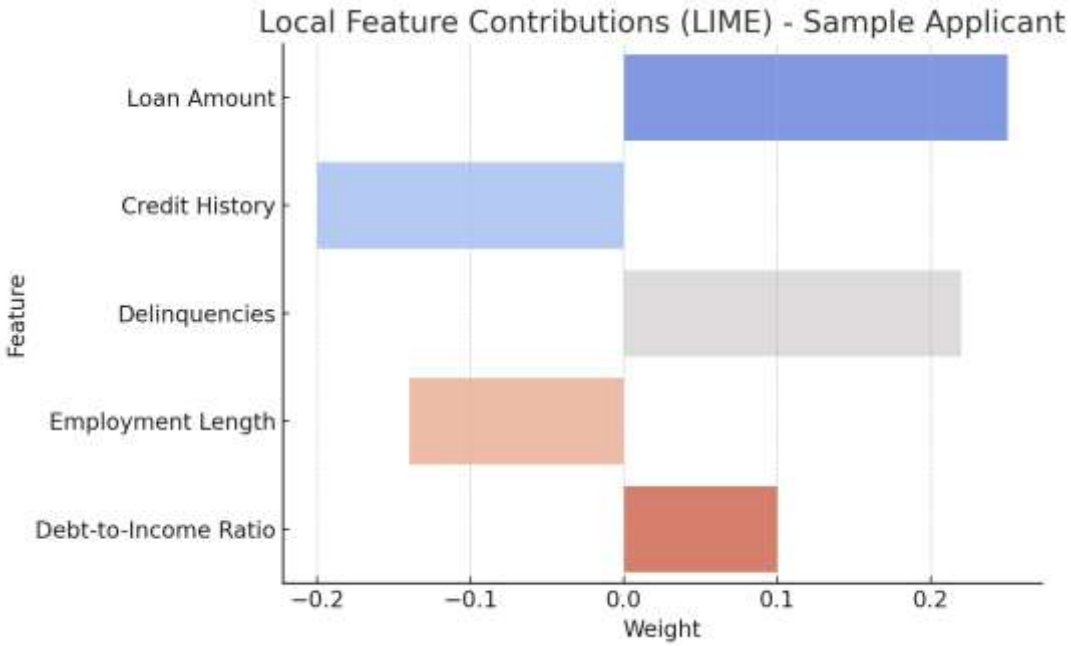

Figure 4: Local Feature Contributions (LIME) - Sample Applicant

Figure 4 displays a bar chart of LIME weights, which represent each feature's influence on the specific prediction. Positive values push the model toward "default," while negative weights support a "non-default" classification. These case-level insights are crucial for human decision-makers such as loan officers. A tabular summary of this LIME explanation is provided in Table 3.

Table 3: LIME Local Explanation Example

Feature	**Weight**
Loan Ammount	0.25
Credit History	-0.2
Delinquencies	0.22
Employment Length	-0.14
Debt-to-Income Ratio	0.1

4.5 Stakeholder Usability and Interpretation

To understand the practical relevance of explainability, we assessed SHAP and LIME outputs from the viewpoint of three stakeholders: loan officers, regulators, and applicants. Loan officers favored SHAP's clear feature rankings for portfolio-level insights, while regulators appreciated its mathematical consistency. On the other hand, individual applicants found LIME's localized breakdown easier to interpret, particularly when evaluating their own creditworthiness. This multi-angle analysis confirms that no single explainability method fits all stakeholders. Effective AI adoption in financial services will require dual deployment: SHAP for compliance and policy-level analysis, and LIME for individualized decisions and appeals.

4.6 Fairness and Ethical Considerations

Beyond performance and usability, ethical AI practice requires fairness evaluation. By analyzing SHAP dependence plots, we observed that some features such as employment title and ZIP code could act as proxies for sensitive attributes like race or socio-economic status. This observation suggests a risk of indirect bias, which may violate fair lending laws and ethical AI standards.
These findings underscore the importance of using XAI not only for interpretability but also as a diagnostic tool for bias detection, feature auditing, and regulatory reporting. Lenders must not only achieve high accuracy but also ensure their models operate in a transparent, fair, and legally compliant manner.

5. Conclusion and Future Work

This study proposed a transparent and interpretable machine learning framework for credit risk assessment by integrating powerful predictive models with state-of-the-art explainable AI techniques. Our experimental analysis demonstrated that while advanced models such as XGBoost and Artificial Neural Networks significantly outperformed traditional logistic regression in predictive accuracy, their opaque nature could hinder trust, compliance, and fairness in financial decision-making. To overcome this limitation, SHAP and LIME were successfully employed to explain both global and local predictions, offering different but complementary insights into the decision-making process. Through SHAP, we uncovered globally influential features like loan-to-income ratio, credit utilization, and delinquency history, all of which are widely accepted indicators in financial risk evaluation. LIME, on the other hand, allowed us to generate localized, instance-specific explanations, which can support case-by-case decision review, particularly beneficial for applicants and loan officers seeking transparency. The results affirmed that a dual-layered explainability strategy global plus local can effectively address the demands of various stakeholders, including financial institutions, regulatory bodies, and customers. Moreover, our fairness assessment revealed that some features may indirectly reflect sensitive personal information, such as ZIP code or employment title, thus highlighting the ethical risks involved in deploying automated systems without interpretability and fairness audits. These findings emphasize the importance of explainable AI not just as a technical solution, but as a mechanism to uphold fairness, accountability, and trust in the financial sector.

Future work can extend this research in several directions. First, incorporating legally sensitive fairness constraints directly into model training may enhance equity in outcomes. Second, longitudinal analysis using time-series features (e.g., behavioral credit scoring over time) could provide deeper insights into creditworthiness dynamics. Finally, user studies involving real-world loan officers and applicants can further validate the practical utility and understandability of XAI tools in operational environments. By

bridging the gap between predictive performance and interpretability, this study contributes to the growing field of responsible AI in finance and lays the groundwork for more transparent, ethical, and effective lending systems.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Thomas, L. C., Crook, J. N., & Edelman, D. B., *Credit Scoring and Its Applications*, SIAM, 2002.
[2] Baesens, B., et al., "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the Operational Research Society*, vol. 54, no. 6, pp. 627–635, 2003.
[3] European Banking Authority, *Guidelines on Loan Origination and Monitoring*, 2020.
[4] Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.
[5] Wachter, S., Mittelstadt, B., & Floridi, L., "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.
[6] Hand, D. J., & Henley, W. E., "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society: Series A*, vol. 160, no. 3, pp. 523–541, 1997.
[7] Hurley, M., & Adebayo, J., "Credit scoring in the era of big data," *Yale Journal of Law & Technology*, vol. 18, pp. 148–216, 2017.
[8] Arrieta, A. B., et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
[9] Galindo, J., & Tamayo, P., "Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications," *Computational Economics*, vol. 15, pp. 107–143, 2000.
[10] Yeh, I. C., & Lien, C. H., "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.
[11] Mehrabi, N., et al., "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.
[12] Van Vlasselaer, V., et al., "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decision Support Systems*, vol. 75, pp. 38–48, 2015.
[13] Ribeiro, M. T., Singh, S., & Guestrin, C., "Why should I trust you?" Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD*, pp. 1135–1144, 2016.
[14] Martens, D., et al., "Comprehensible credit scoring models using rule extraction from support vector machines," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1466–1476, 2007.
[15] Chen, T., & Guestrin, C., "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD*, pp. 785–794, 2016.
[16] Goodfellow, I., Bengio, Y., & Courville, A., *Deep Learning*, MIT Press, 2016.
[17] Wang, H., et al., "A deep learning approach for credit risk evaluation," *Applied Intelligence*, vol. 49, pp. 315–328, 2019.
[18] Ribeiro, M. T., Singh, S., & Guestrin, C., "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.
[19] Lundberg, S. M., & Lee, S. I., "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.
[20] Martens, D., & Provost, F., "Explaining data-driven document classifications," *MIS Quarterly*, vol. 38, no. 1, pp. 73–99, 2014.
[21] Bhatia, M., & Aggarwal, S., "Credit risk prediction with model interpretability using SHAP values," *Journal of Banking and Financial Technology*, vol. 5, pp. 1–12, 2021.
[22] Caruana, R., et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proceedings of the 21st ACM SIGKDD*, pp. 1721–1730, 2015.
[23] Brownlee, J., *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*, Machine Learning Mastery, 2020.
[24] Guyon, I., et al., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.
[25] Chawla, N. V., et al., "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[26] He, H., & Garcia, E. A., "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
[27] Hall, P., Gill, N., & Kurka, M., "Machine learning interpretability: A science rather than a tool," *arXiv preprint arXiv:1902.03855*, 2019.
[28] Lipton, Z. C., "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.
[29] MLflow Documentation, "Open Source Platform for the Machine Learning Lifecycle," https://mlflow.org/, Accessed May 2025.
[30] Fernández-Delgado, M., et al., "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.
[31] Štrumbelj, E., & Kononenko, I., "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.
[32] Arya, V., et al., "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.
[33] Doshi-Velez, F., & Kim, B., "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[34] Barocas, S., Hardt, M., & Narayanan, A., *Fairness and Machine Learning: Limitations and Opportunities*, fairmlbook.org, 2021.

[35] Lepri, B., et al., "Fair, transparent, and accountable algorithmic decision-making processes," *Philosophy & Technology*, vol. 31, no. 4, pp. 611–627, 2018.

[36] Md Sohanur Rahman Sourav, Arafat Hossain, Md Redwanul Islam, Mohtasim Wasif, & Sujana Samia. (2025). AI-Driven forecasting in BRICS infrastructure investment: impacts on resource allocation and project delivery. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 117-132. https://doi.org/10.32996/jefas.2025.7.2.11

[37] Mohtasim Wasif, Sujana Samia, Md Sohanur Rahman Sourav, Arafat Hossain, & Md Redwanul Islam. (2025). Data-Driven insights on the relationship between BRICS financial policies and global investment trends. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 133-147. https://doi.org/10.32996/jefas.2025.7.2.12

[38] Md Redwanul Islam, Mohtasim Wasif, Sujana Samia, Md Sohanur Rahman Sourav, & Arafat Hossain. (2025). The Role of Machine Learning in Forecasting U.S. GDP Growth after the COVID-19 Pandemic. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 163-175. https://doi.org/10.32996/jefas.2025.7.2.14

[39] Md. Tanvir Rahman Mazumder, Md. Shahadat Hossain Shourov, Iftekhar Rasul, Sonia Akter, & Md Kauser Miah. (2025). Fraud Detection in Financial Transactions: A Unified Deep Learning Approach. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 184-194. https://doi.org/10.32996/jefas.2025.7.2.16

[40] Md. Tanvir Rahman Mazumder, Md. Shahadat Hossain Shourov, Iftekhar Rasul, Sonia Akter, & Md Kauser Miah. (2025). The Impact of Macroeconomic Factors on the U.S. Market: A Data Science Perspective. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 208-219. https://doi.org/10.32996/jefas.2025.7.2.18

[41] Md. Tanvir Rahman Mazumder, Md. Shahadat Hossain Shourov, Iftekhar Rasul, Sonia Akter, & Md Kauser Miah. (2025). Anomaly Detection in Financial Transactions Using Convolutional Neural Networks. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 195-207. https://doi.org/10.32996/jefas.2025.7.2.17

[42] Newaz, A. A. H., Mitra, R., Jahan, R., & Kadir, A. (2025). Free Vibration Characteristics of Single-Degree-of-Freedom (SDOF) Mechanical Systems: Investigating through Theory, Experimentation and Numerical Simulation. *Engineering Research: Perspectives on Recent Advances Vol. 7*, 43–57. https://doi.org/10.9734/bpi/erpra/v7/5456

[43] Abdullah Al Hossain Newaz, Kazi Abdullah Al Imon, Refat Jahan, and Imran Khan Tanvir. 2022. "Advanced Motor Design and Optimization for High-Efficiency Industrial Applications". Metallurgical and Materials Engineering 28 (4):697-713. https://doi.org/10.63278/mme.v28i4.1282.

[43] Comprehensive Dynamic Modeling of a Rotary Servo Base Unit Using Frequency Response and Bump Test Techniques
American Journal of Mechanical Engineering. 2025, 13(1), 6-10
DOI: https://pubs.sciepub.com/ajme/13/1/2/index.html

[44] Revolutionizing American Military Protection: Development and Implementation of Next-Generation Shielding System
North American Academic Research. (2025). Revolutionizing American Military Protection: Development and Implementation of Next-Generation Shielding System. In North American Academic Research (Vol. 8, Number 1). Zenodo. https://doi.org/10.5281/zenodo.14927622

[45]Lean Six Sigma Implementation of USA Military North American Academic Research. (2025). Lean Six Sigma Implementation of USA Military. In North American Academic Research (Vol. 8, Number 1). Zenodo. https://doi.org/10.5281/zenodo.14927559

[46] Md Sohanur Rahman Sourav, Arafat Hossain, Md Redwanul Islam, Mohtasim Wasif, & Sujana Samia. (2025). AI-Driven forecasting in BRICS infrastructure investment: impacts on resource allocation and project delivery. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 117-132. https://doi.org/10.32996/jefas.2025.7.2.11

[47] Mohtasim Wasif, Sujana Samia, Md Sohanur Rahman Sourav, Arafat Hossain, & Md Redwanul Islam. (2025). Data-Driven insights on the relationship between BRICS financial policies and global investment trends. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 133-147. https://doi.org/10.32996/jefas.2025.7.2.12

[48] Md Redwanul Islam, Mohtasim Wasif, Sujana Samia, Md Sohanur Rahman Sourav, & Arafat Hossain. (2025). The Role of Machine Learning in Forecasting U.S. GDP Growth after the COVID-19 Pandemic. *Journal of Economics, Finance and Accounting Studies* , *7*(2), 163-175. https://doi.org/10.32996/jefas.2025.7.2.14

[49] Abir, S. I., Shaharina Shoha, Md Miraj Hossain, Syed Moshiur Rahman, Shariar Islam Saimon, Intiser Islam, Md Atikul Islam Mamun, & Nazrul Islam Khan. (2024). Deep Learning-Based Classification of Skin Lesions: Enhancing Melanoma Detection through Automated Preprocessing and Data Augmentation. *Journal of Computer Science and Technology Studies*, *6*(5), 152-167. https://doi.org/10.32996/jcsts.2024.6.5.13

[50] Abir, S. I., Shaharina Shoha, Sarder Abdulla Al Shiam, Shariar Islam Saimon, Intiser Islam, Md Atikul Islam Mamun, Md Miraj Hossain, Syed Moshiur Rahman, & Nazrul Islam Khan. (2024). Precision Lesion Analysis and Classification in Dermatological Imaging through Advanced Convolutional Architectures. *Journal of Computer Science and Technology Studies*, *6*(5), 168-180. https://doi.org/10.32996/jcsts.2024.6.5.14

[51] Nigar Sultana, Shariar Islam Saimon, Intiser Islam, Abir, S. I., Md Sanjit Hossain, Sarder Abdulla Al Shiam, & Nazrul Islam Khan. (2025). Artificial Intelligence in Multi-Disease Medical Diagnostics: An Integrative Approach. *Journal of Computer Science and Technology Studies*, *7*(1), 157-175. https://doi.org/10.32996/jcsts.2025.7.1.12

[52] Abir, S. I., Shaharina Shoha, Sarder Abdulla Al shiam, Nazrul Islam Khan, Abid Hasan Shimanto, Muhammad Zakaria, & S M Shamsul Arefeen. (2024). Deep Learning Application of LSTM(P) to predict the risk factors of etiology cardiovascular disease. *Journal of Computer Science and Technology Studies*, *6*(5), 181-200. https://doi.org/10.32996/jcsts.2024.6.5.15