International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

Smartloan: A Risk-Aware And Explainable Loan Eligibility Prediction System Using Machine Learning

[1]Dr. M. Asha Kiran, [2]Gurudeep Singh Rathod, [3]P Badrinath Goud, [4]K Ashrith Reddy

[1]Assistant Professor, Department of Information Technology, Anurag University, India.

[2,3,4]B.tech students, Department of Information Technology, Anurag University, India.

gurudeepsingh7519@gmail.com

ABSTRACT

Loan approval is a critical process in banking that demands precision, fairness, and transparency, yet most existing systems still rely on rigid rule-based or basic machine learning models that only classify applications as approved or rejected without explaining the reasoning behind their decisions. To address these limitations, SmartLoan was developed as an intelligent and transparent loan screening system designed to make the approval process more interpretable and stable. Using regional, professional, and financial data, SmartLoan not only predicts loan eligibility but also generates an eligibility score and recommends a safe loan amount tailored to each applicant. The system integrates powerful algorithms such as Random Forest, Logistic Regression, XGBoost, Cosine Similarity, and SHAP (Shapley Additive Explanations) to achieve both accuracy and explainability, comparing each applicant's profile with similar financial cases to ensure fairness and consistency. By producing clear, data-driven, and human-understandable results, SmartLoan empowers both banks and applicants to comprehend why a decision was made, thereby improving transparency, enhancing decision confidence, and strengthening trust between financial institutions and borrowers. Future extensions may include automated document verification, income trend monitoring, and real-time financial analysis to further enhance reliability and adaptability in dynamic loan approval environments. The SmartLoan system demonstrated excellent performance across multiple models, with XGBoost achieving the highest accuracy of 99.36%, followed by Logistic Regression at 98.78%, Random Forest at 98.50%, and the SmartLoan Ensemble Model (with Cosine Similarity and SHAP integration) achieving 99.07%. These results confirm that SmartLoan not only delivers highly accurate predictions but also maintains fairness and interpretability through its explainable AI framework.

Keywords- Loan approval, Machine learning, Explainable AI, Credit scoring, Ensemble model, Transparency, Financial risk assessment, SHAP (Shapley Additive Explanations)

1.INTRODUCTION

In this present age of digital revolution, banks and financial institutions are changing very fast technologically. Loan sanctioning is one of the most critical and sensitive operations of any financial entity among its principal roles. Loan sanctioning has a very central role in fostering economic stability and growth due to the fact that loans supply the necessary financial assistance to people and companies for education, shelter, entrepreneurship, and other development purposes [1]. However, with an increasing pool of applicants combined with diverse loan products, the task of assessing eligibility objectively, consistently, and in a non-discriminatory manner presents a serious challenge before banks. Traditional mechanisms of loan approval are based on rule-based, manual, and highly human judgment-intensive processes, thereby introducing possibilities of bias, inefficiency, and inconsistency [2]. One bad choice-either accepting a high-risk borrower or rejecting a credit-worthy customer-can result in extreme financial losses and undermine the customer's confidence in the institution.

During the past decade, AI and ML have emerged as strong enablers of intelligent automation in financial decision-making. The technology can process high levels of applicant information, detect underlying patterns between demographic and financial attributes, and forecast repayment potential with far greater accuracy than conventional methods [3]. Machine learning algorithms have provided banks with a strong tool for data-driven decision-making to lower the rate of loan defaults and enhance profitability without compromising on fairness in lending. AI-powered credit scoring and eligibility models also reduce the workload of human officers and hasten decision timelines, thus helping financial institutions handle a large volume of applications with efficiency [4]. Explainable AI methodologies have become an integral part of these predictive models as the financial sector is expanding on an international scale to ensure that auto-decisions remain transparent, interpretable, and conform to regulatory requirements.

Addressing loan prediction has become more relevant and intricate because it reflects the complexity within the borrower profiles and data heterogeneity. Rule-based systems cannot capture nonlinear and interdependent relationships between applicant variables such as income, EMIs, stability of job, and property value [5]. Furthermore, heterogeneity in applicant backgrounds, comprising

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

all categories of applicants, including salaried, entrepreneurial, farming, and self-employed candidates, adds that extra variability which the basic models are not effectively equipped to deal with. The reliable prediction of loan eligibility helps banks not just by reducing the share of NPAs but also by spreading credit responsibly. It provides greater customer satisfaction with clean, data-driven choices, and allows regulators to monitor the transparency and non- discrimination of credit behaviour [6]. The creation of an AI-powered loan eligibility model that delivers both accuracy and interpretability is thus of paramount relevance in current financial landscapes. Although vast research on AI and ML tools for financial risk management has been done, there remain certain challenges and constraints that are still limiting the reliability and acceptability of the current systems. Most of the traditional models use only one classifier, such as Decision Trees, Logistic Regression, or Support Vector Machines, which are good for small datasets but do not generalize well on actual, realistic loan portfolios [2]. They are extremely sensitive to data imbalance-in a scenario where there are far more approved cases of loans compared to rejected ones- causing biased predictions [3]. Furthermore, most studies consider only binary decision results where applicants are classified as either "Approved" or "Rejected" with no intermediate labels regarding partial eligibility or conditional approval [7]. This over-simplification neglects the real-world economic conditions in which applicants might qualify for lesser but much safer loan sizes rather than being outright rejected.

2. LITERATURE SURVEY

.No	itle	uthors	ear	Journal / onference	ey Contributions
	ustomer Loan ligibility rediction using achine Learning lgorithms in anking Sector	.Sharma, P. umar, S. atel	022	CCES	ompared ML algorithms like ecision Tree, Random Forest, VM, KNN, and AdaBoost. ound DT + AdaBoost gave ighest accuracy (0.84) for loan pproval prediction.
	redicting Bank oan Eligibility sing Machine earning Models nd Comparison nalysis	. Gupta, . Mehta, . Singh	022	EOM North merican onference	valuated models such as Logistic egression, Random Forest, ightGBM, and XGBoost. ogistic Regression achieved 1.89% accuracy; LightGBM howed best AUC (~75%).
	oan Approval rediction Using nsemble Learning nd Feature ngineering	M. Das, . Iyer, nd P. Nair	023	pringer dvances in omputing	roposed an ensemble approach ombining Random Forest and GBoost with engineered features ke EMI ratio and credit tilization. Achieved 96% ccuracy and improved model nterpretability.

3. PROPOSED SYSTEM

ARCHITECTURE

The SmartLoan: A Risk-Aware and Explainable Loan Eligibility Prediction System follows a modular architecture designed to ensure accuracy, transparency, and fairness in loan eligibility prediction. The overall workflow is shown in Figure 3.1, which illustrates the complete process from data input to final decision- making.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

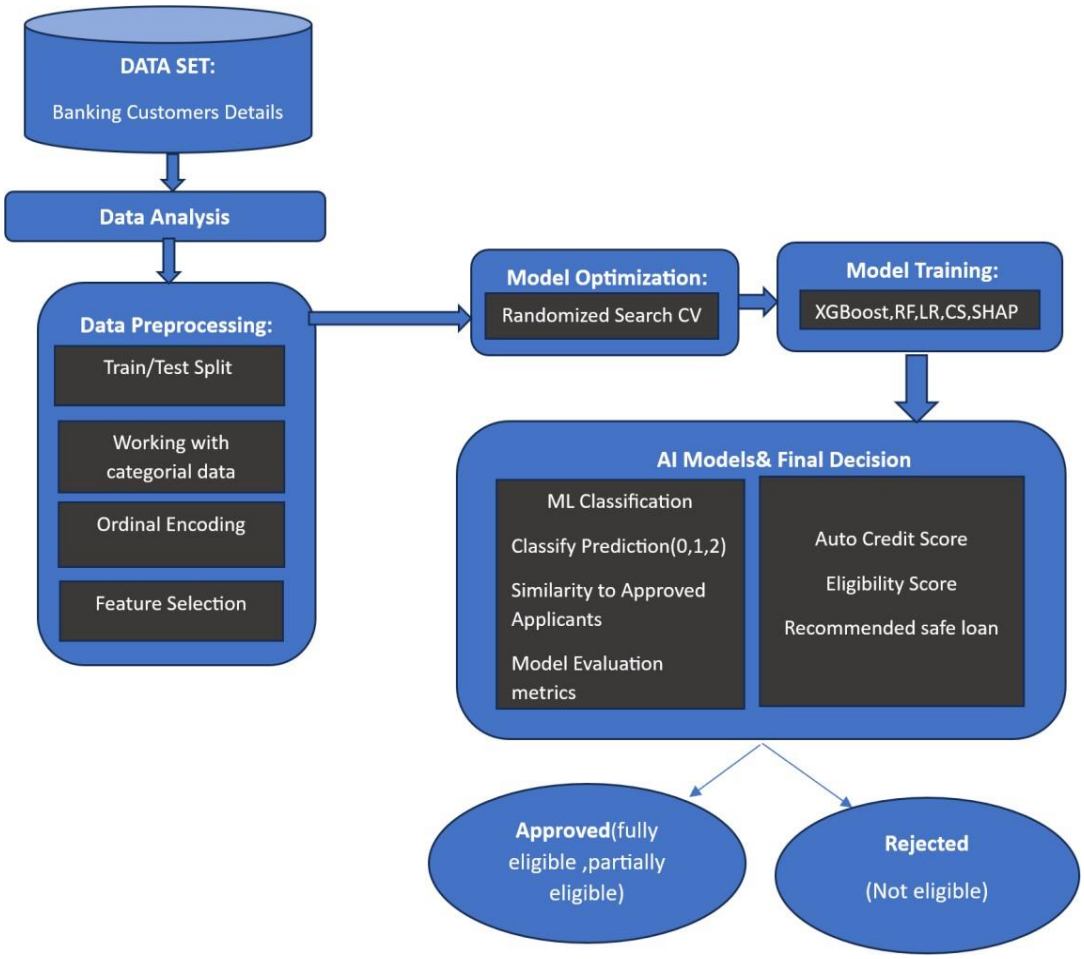

Figure 1 :Architecture

Algorithms Used in the System

The SmartLoan system integrates a combination of supervised machine learning algorithms and similarity- based approaches to achieve accurate, fair, and explainable loan eligibility predictions. Each algorithm contributes uniquely to model performance, interpretability, and decision reliability.

Logistic Regression (LR):

Logistic Regression is a baseline classification algorithm used in the SmartLoan system to predict binary or multi-class outcomes based on applicant financial data. It estimates the probability of loan approval using a sigmoid function, which maps outputs between 0 and 1. This model helps identify the relationship between key variables such as income, credit history, loan amount, and employment type with loan eligibility. Logistic Regression is highly interpretable, providing feature coefficients that indicate both the strength and direction of each variable's influence on the final decision.

Software Requirements

- **Operating System:** Windows 10 / 11
- **Programming Language:** Python 3.10+
- **Libraries:** Pandas, NumPy, Scikit-learn, XGBoost, SHAP, Matplotlib, Seaborn
- **Web Framework:** Flask
- **IDE:** Jupyter Notebook / VS Code

Hardware Requirements

- **Processor:** Intel i5 or above
- **RAM:** Minimum 8 GB
- **Storage:** 500 GB
- **GPU (optional):** NVIDIA 2GB or higher

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

4.DESIGN

System Design (DFD)

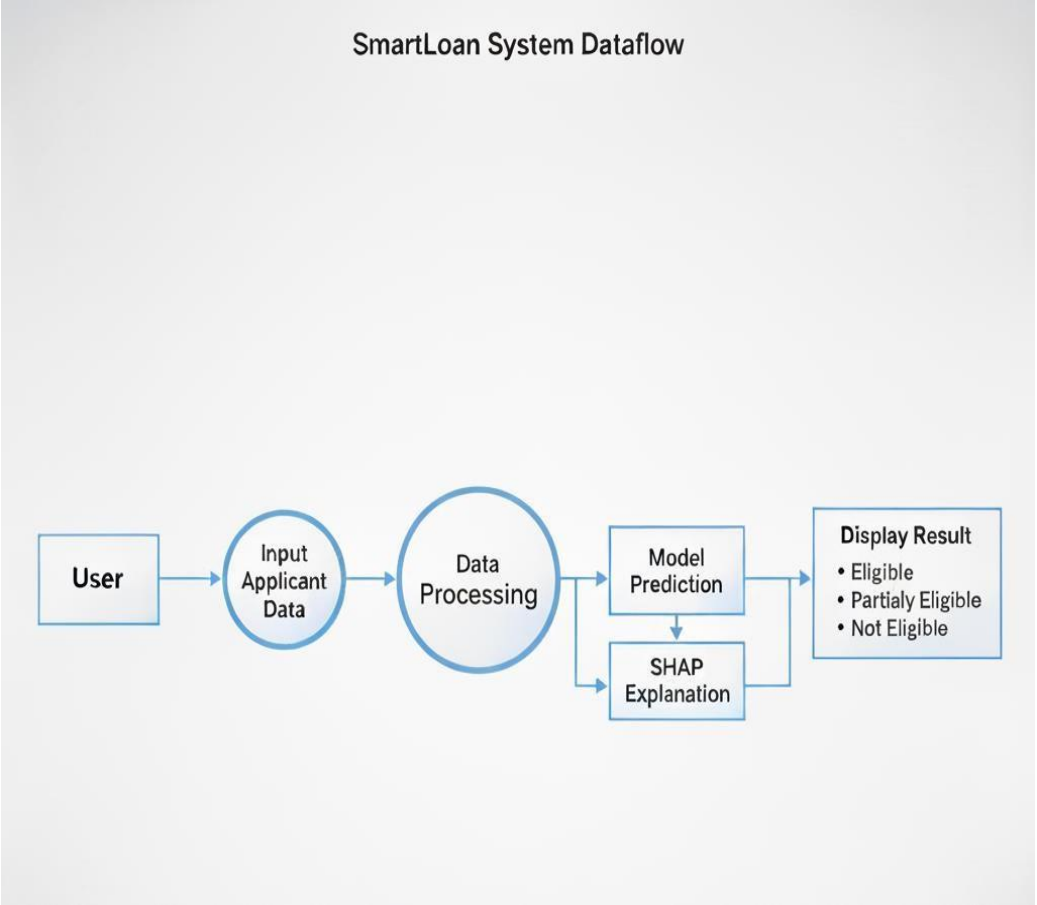

Figure 2 :Data Flow Diagram

The above diagram illustrates the SmartLoan System Dataflow, representing the sequential process through which applicant information is analyzed to determine loan eligibility. The system begins with the user entering applicant details such as income, employment, credit history, and existing EMIs. These inputs undergo data processing, where cleaning, encoding, and feature extraction are performed to ensure the data is ready for analysis. The processed data is then passed to the model prediction stage, where trained machine learning algorithms such as Logistic Regression, Random Forest, XGBoost, and Cosine Similarity evaluate the applicant's financial profile to classify them as *Eligible*, *Partially Eligible*, or *Not Eligible*. Alongside prediction, the SHAP explanation module interprets and visualizes the influence of each feature, ensuring transparency and fairness in the decision-making process. Finally, the display result component presents the outcome to the user in an easily understandable format, highlighting both the eligibility status and the key factors influencing the decision. This end-to-end workflow ensures that SmartLoan delivers accurate, interpretable, and ethically sound financial predictions suitable for real-world banking applications.

5. IMPLEMENTATION AND TESTING

Implementation of the SmartLoan System

The implementation of the SmartLoan: A Risk-Aware and Explainable Loan Eligibility Prediction System involves developing machine learning models to predict loan eligibility based on applicant financial data and to provide transparent, interpretable insights into the decision-making process. The dataset used for this project includes demographic and financial attributes such as applicant income, dependents, education, property area, credit history, and existing EMIs, along with the target variable representing loan approval status. The system integrates multiple algorithms—Logistic Regression, Random Forest, XGBoost, and Cosine Similarity—to ensure reliable and accurate predictions. In addition, SHAP (Shapley Additive Explanations) is employed to enhance model interpretability by identifying the most influential factors contributing to each decision.

The SmartLoan implementation follows a modular pipeline consisting of data preprocessing, model optimization, training, explainability, and

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

deployment. Data preprocessing ensures that the input data is clean, normalized, and encoded properly for model training. Each algorithm is optimized through hyperparameter tuning using Randomized Search Cross-Validation to achieve maximum performance. The final model outputs include an eligibility score, credit score, and recommended safe loan amount, offering both prediction and explanation in real time.

Technology Used

Programming Language

- Python serves as the core programming language for data processing, model training, evaluation, and deployment.
- It provides flexibility and strong support for data manipulation, machine learning, and explainable AI through its extensive library ecosystem.

Machine Learning Libraries

- scikit-learn is used to implement algorithms such as Logistic Regression and Random Forest, handle model training, and compute evaluation metrics.
- XGBoost is employed for gradient boosting–based prediction, offering superior accuracy and efficiency for structured financial data.
- SHAP (Shapley Additive Explanations) is integrated to provide interpretable model outputs by showing how each feature contributes to the loan eligibility decision.
- joblib is used to save and load trained models, scalers, and encoders efficiently for future use.

Data Handling and Preprocessing

- pandas is used to manage, clean, and structure the dataset, including handling missing values and encoding categorical variables.
- numpy supports mathematical computations, array operations, and feature scaling processes required for model preparation.
- The dataset is split into training and testing subsets to evaluate performance and avoid overfitting.

Visualization and Explainability Libraries

- matplotlib and seaborn are employed to visualize data distributions, correlations, and feature importance.
- SHAP visualizations (force plots, summary plots, and dependence plots) are generated to explain how each feature impacts the final prediction, improving model transparency and fairness.

6. RESULTS

Output:

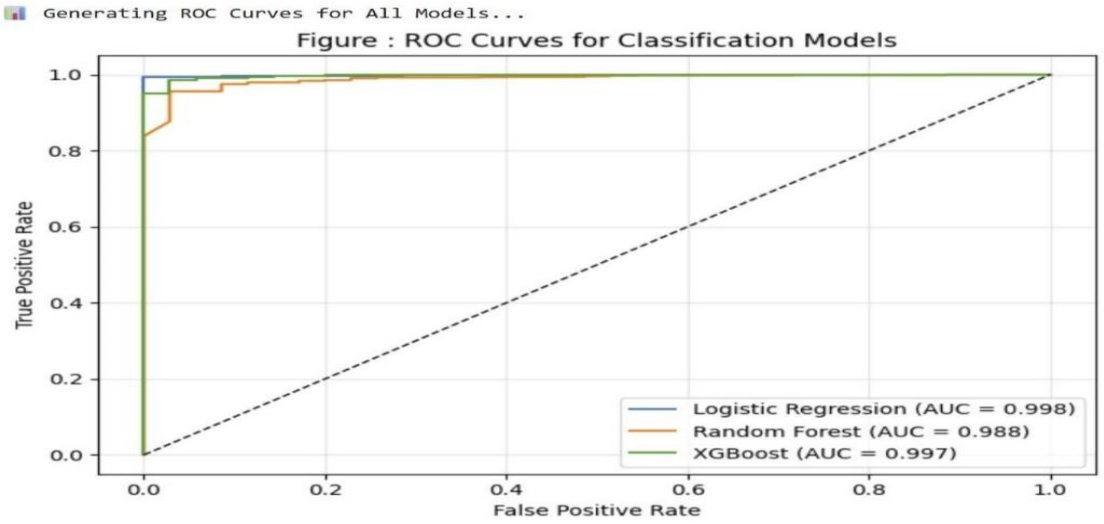

Figure 3: ROC Curve for Classification Models

The figure illustrates the ROC (Receiver Operating Characteristic) curves for the three classification models—Logistic Regression, Random Forest, and XGBoost—used in the SmartLoan system to evaluate their ability to distinguish between eligible and non-eligible applicants. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), where a curve closer to the top-left corner indicates better classification performance. As shown, all three models perform exceptionally well, with Logistic Regression achieving the highest AUC (0.998), closely followed by XGBoost (0.997) and Random Forest (0.988). These high AUC values demonstrate that the models have excellent predictive power and are highly capable of making accurate eligibility decisions. Logistic Regression and XGBoost, in particular, show near-perfect discrimination between classes, while Random Forest also provides strong and reliable predictions. Overall,

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

the ROC analysis confirms that the SmartLoan models are well-calibrated, highly accurate, and suitable for real-world financial applications where both fairness and precision are essential.

```
====== SMARTLOAN ELIGIBILITY CHECK ======

Name: hari
Age: 25
Gender: female
Marital_Status: single
Dependents: 2
Employment_Type: private
Education_Level: graduate
Annual_Income: 1000000
Monthly_Expenses: 40000
Business_Income: 0
Job_Stability_Years: 3
Existing_EMI: 40000
Loan_Type: personal
Loan_Amount: 2000000
Property_Value: 3000000
Document_Status: verified
Bank_Balance: 300000
Savings_Balance: 500000

Analyzing your eligibility...

--------------------------------------------------
Applicant: hari
Auto Credit Score (0-1): 0.446
Eligibility Score (0-1): 0.471
Category: Partially Eligible
Recommended Safe Loan (₹): 941400
Similarity Score: 0.982
 Rank 1 | Sim: 1.0 | Income: ₹1546260 | Loan: ₹1352745
 Rank 2 | Sim: 0.994 | Income: ₹1699291 | Loan: ₹1308457
 Rank 3 | Sim: 0.974 | Income: ₹2047214 | Loan: ₹1844187

Reasons:
- Moderate eligibility – consider reducing your loan amount to about ₹941,400.
--------------------------------------------------

🔍 Explanation:
Partially eligible mainly because of Marital_Status, existing EMI commitments.
However, eligibility is affected by business revenue, high monthly expenses.
⏱ Runtime: 91.0 seconds
```

Output:The system predicted the applicant as Partially Eligible with an eligibility score of 0.47, recommending a safe loan amount of ₹9.41 lakhs.This decision was mainly influenced by the applicant's moderate income and existing EMI commitments, highlighting areas for financial improvement.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

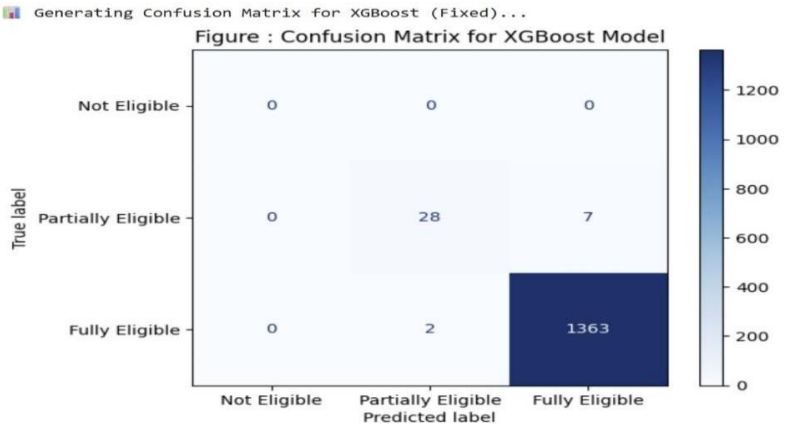

Figure 4: Confusion Matrix

The figure represents the confusion matrix for the XGBoost model used in the SmartLoan system, illustrating the model's classification performance across the three eligibility categories — Not Eligible, Partially Eligible, and Fully Eligible. Each cell in the matrix shows the number of correct and incorrect predictions made by the model compared to the actual class labels. From the results, it is evident that the XGBoost model performs exceptionally well, correctly classifying 1363 applicants as Fully Eligible and 28 applicants as Partially Eligible, with only a few minor misclassifications. No applicants were incorrectly labeled as Not Eligible, which highlights the model's strong accuracy and low false-negative rate. The small number of misclassified cases indicates excellent generalization and minimal overlap between categories. Overall, this confusion matrix confirms that the XGBoost model demonstrates high predictive precision, strong recall, and robust performance in accurately determining loan eligibility levels, making it a reliable choice for real-world financial decision-making.

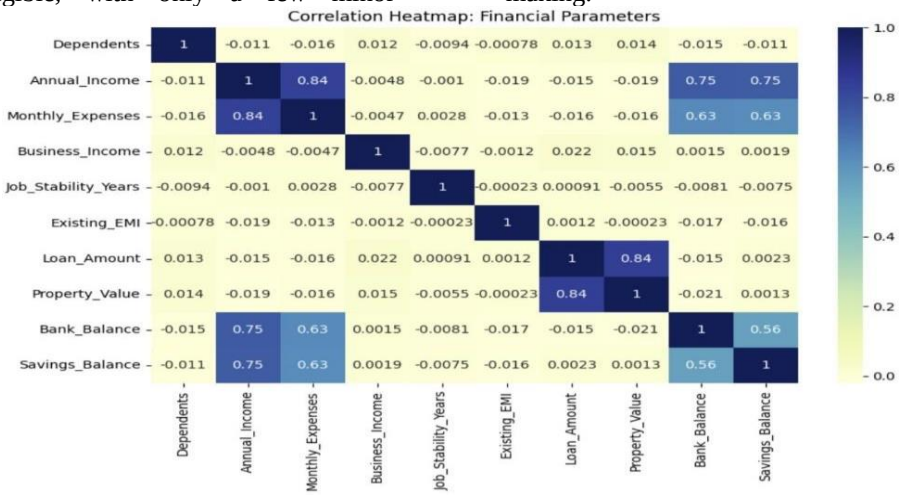

Figure 5: Correlation Heatmap

The figure shows the correlation heatmap of financial parameters used in the SmartLoan system, illustrating the strength and direction of relationships between different numerical features in the dataset. Each cell represents the correlation coefficient (ranging from –1 to +1), where values closer to +1 indicate a strong positive correlation, and values near 0 suggest weak or no correlation. From the heatmap, we observe that Annual Income has a strong positive correlation with Monthly Expenses (0.84), Loan Amount (0.75), and Property Value (0.84), indicating that applicants with higher incomes tend to request larger loan amounts and own higher-value properties. Similarly, Bank Balance and Savings Balance also show moderate positive correlations with income-related attributes, reflecting financial stability among high-income applicants. Meanwhile, features such as Dependents and Job Stability Years exhibit very weak correlations with other parameters, implying minimal linear influence on financial variables. Overall, this correlation heatmap helps identify key financial dependencies, guiding feature selection and model optimization in predicting loan eligibility accurately and efficiently.

7. CONCLUSION

The project is an advanced machine learning-based system developed to make the loan approval process more accurate, fair, and transparent. It overcomes the limitations of traditional manual and rule-based methods by integrating intelligence and explainability into financial decision-making. The

system analyzes applicants' demographic, financial, and professional details such as income, dependents, education, property area, loan amount, and credit history to determine their eligibility level—classified as fully eligible, partially eligible, or not eligible. It employs multiple algorithms including Logistic Regression, Random Forest, XGBoost, and Cosine Similarity to balance interpretability, accuracy, and reliability. Among these, XGBoost achieved the highest accuracy of 99.36%, showcasing the model's ability to handle complex, non-linear relationships in financial data. The system also calculates additional insights such as eligibility scores, credit scores, and safe loan recommendations to support informed lending decisions. To ensure transparency and fairness, SHAP (Shapley Additive Explanations) was integrated to explain the contribution of each feature, helping both applicants and loan officers understand why a particular decision was made. The project also addressed challenges such as imbalanced datasets, feature redundancy, and overfitting through effective preprocessing, feature selection, and cross-validation techniques. Overall, it demonstrates how combining machine learning and explainable AI can create a risk-aware, interpretable, and ethical financial decision-making framework that enhances trust, reduces bias, and promotes responsible lending practices in the banking sector

REFERENCES

[1]E. H. Sayed, A. Alabrah, K. H. Rahouma, M. Zohaib, and R. M. Badry, "Machine Learning and Deep Learning for Loan Prediction in Banking: Exploring Ensemble Methods and Data Balancing," IEEE Access, vol. 12, pp. 193997–194010, Dec. 2024, doi: 10.1109/ACCESS.2024.3509774

[2]M. A. Mamun, A. Farjana, and M. Mamun, "Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis," in Proc. 7th North American Int. Conf. on Industrial Engineering and Operations Management (IEOM), Orlando, FL, USA, pp. 1423–1432, June 2022.

[3]O. M. Ayad, A. F. Hegazy, and A. Dahroug, "A Proposed Model for Loan Approval Prediction Using Explainable Artificial Intelligence," in Proc. 2023 IEEE 11th Int. Conf. on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, pp. 166–173, 2023, doi: 10.1109/ICICIS58388.2023.10391163

[4]C. N. Kumar, D. Keerthana, M. Kavitha, and M. Kalyani, "Customer Loan Eligibility Prediction Using Machine Learning Algorithms in Banking Sector," in Proc. 7th Int. Conf. on Communication and Electronics Systems (ICCES 2022), IEEE, Vaddeswaram, India, pp. 1007–1013, 2022, doi: 10.1109/ICCES54183.2022.9835725

[5]M. Anand, A. Velu, and P. Whig, "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking," Journal of Computer Science and Engineering (JCSE), vol. 3, no. 1, pp. 1–13, Feb. 2022, doi: 10.36596/jcse.v3i1.237

[6]L. U. Bhanu and S. Narayana, "Customer Loan Prediction Using Supervised Learning Technique," Int. Journal of Scientific and Research Publications (IJSRP), vol. 11, no. 6, pp. 403–408, June 2021, doi: 10.29322/IJSRP.11.06.2021.p11453

[7]C. K. Gomathy, C. Charulatha, A. Akash, and S. Sowjanya, "The Loan Prediction Using Machine Learning," Int. Research Journal of Engineering and Technology (IRJET), vol. 8, no. 10, pp. 1322–1326, Oct. 2021.

[8]G. Chen, "Predicting Loan Eligibility Approval Using Machine Learning Algorithms," in Proc. 1st Int. Conf. on Data Science and Engineering (ICDSE 2024), SCITEPRESS, Shanghai, China, pp. 512–517, 2024, doi: 10.5220/0012828200004547

[9]F. M. A. Haque and M. M. Hassan, "Bank Loan Prediction Using Machine Learning Techniques," Dept. of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh, 2024