iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68388

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

TRUSTNOVA: A Predictive Machine Learning Model for Loan Approval

Krutika Kolpate[1], Antariksha Dhanure[2], Gargi Fadtare[3], Gayatri Gunjal[4], Dr. Nilesh Bhelkar[5]
Artificial Intelligence And Data Science, Rajiv Gandhi Institute of Technology, Mumbai

Abstract: ***Small low-income banks have problems in making accurate decisions in the approval of loans for low-income earners. Manual checking tends to make there be inefficiency bias and loss of opportunities. This study designs a Loan Approval Prediction System empowered with machine learning that automates the process and optimizes the processes involved in the assessment of eligibility for a loan. Data analysis on applicant details — income, credit history, and repayment capacity — would bring fairer, faster, and more accurate results in the lending decisions for banks. The model provides banks with the necessary conditions for making lending decisions that are fair and efficient. This will in the long run improve the financial inclusion of the unbanked and underbanked populations.***

Keywords: Loan Approval, Machine Learning, Predictive Modeling, Financial Inclusion, Risk Assessment, Credit Scoring, Explainable AI, Alternative Credit Data, Bias

I. INTRODUCTION

The approval of loans is a big factor in the encouragement of financial inclusion; this is how the risks involved can be assessed. So, credit scoring has to be done. This is one of the very core functions for banks serving low-income earners, and more so, the microfinance banks. Manual review is long and subjective. There is a high possibility of the conscious or unconscious bias loan officers can introduce into the system, and any subjective judgment meted out to borrowers might lead to an inequality exposure channel toward the application. ML Predictive model enhances the formation of sound underwriting and credit decisions that will need to be both accurate and perfectly unbiased. This paper traces the development of a system for predicting loan approval using classification algorithms to appraise the probability of approval based on applicant data with the general, dual purpose of increasing operational efficiency and improving fairness. The importance of financial inclusion can hardly be overemphasized since credit access enables individuals to invest in business, education, and housing. The application of ML models will enable banks to reach more of the market and cater to a much wider client base, including those without much credit history. This paper will further detail the ethical requirements and the necessity for maintaining a proportionate level of human control over automation to avoid algorithmic discrimination. A multitude of studies have proved that ML indeed works best for financial decisions. Logistic regression or even decision trees and ensemble methods, namely random forests, are indeed used in credit risk, as many papers cited proof. The models will have overfitting problems or bias issues if not treated properly. This finding also comes into view with regard to feature selection, where credit score, income, and loan amount become very important in predicting the outcome of the loan. Recent strides in the explainable AI (XAI) front, SHAP values, and LIME, to be precise have led to the possibility of explaining model decisions and, in the long run, to building trust in automated systems. Meanwhile, studies confirm that integrating non-traditional data sources — utility bill payments and mobile phone usage and social behavior — increases predicting efficiency regarding underserved populations. For example, evidence has it that constant rent payments or periods of regular mobile top-ups could relate impliedly to financial loyalties even within a missing formal credit score.

II. LITERATURE REVIEW

The machine learning methods which can be applied in the prediction of the probability of approval granted by the bank for the loan as well as to identify which icons are worthy of approval include; Random Forest, DT, NB, Logistic Regression. In order to anticipate loan eligibility and give users a quick loan status check, this study [1] presents an intelligent loan helper that uses machine learning techniques and achieved 88% accuracy. In [2] for banks and consumers alike, machine learning techniques—Random Forest in particular—have the ability to anticipate loan acceptance and affordability with high accuracy, streamlining the approval process. [3] By offering a more accurate and efficient substitute for conventional techniques, machine learning algorithms can completely transform the loan approval process by lowering errors and enhancing decision-making. In [4] Banks can use machine learning algorithms to assess a customer's creditworthiness, which helps with loan request approval or rejection.

The author implement SVM, DT and XGboost ML algorithms and achieved 81% accuracy. The author [5] used the Random Forest algorithm, machine learning models can increase the efficiency, speed, and accuracy of bank loan approval processes. In [6] machine learning models trained on historical data can forecast whether a new loan application will be approved or denied, thus streamlining the loan approval process in the banking industry. The author [7] examining loan payback likelihood and client credit history, Support Vector Machine (SVM) and Random Forest (RF) algorithms forecast loan acceptance for customers with high accuracy [8]. By efficiently using machine learning algorithms, enhancing accuracy and performance is still a challenge [9]. To avoid spending considerable amount of time on manually going through the applications to approve, loan banks have been developed to help in prioritizing them. It will also improve satisfaction of customers since the applications are processed within the agreed time. In general, this project is intended to enhance the efficiency of loan issuance and increase the probability of lending success for both banks and applicants; data mining techniques such as clustering, association, and classification can determine whether a loan application will be approved or rejected, helping the banks makes their decision without risk of loss.[10] Other methods of machine learning involve the ability of the random forest method, which can identify probable defaulters from customers' loan-approved history with a higher precision compared with other methods [11]. Based on the needs of the banking industry to assess loan approval risk, this study proposes an adaptive machine learning model for automatic loan approval that utilizes four categorization algorithms [12]. While as by comparing the results of the model with other machine learning algorithms then it has found that Logistic Regression has the highest accuracy of 92% and F1 Score of 96% for successfully classifying bank loan eligibility [13]. It tries to help financial institutions reduce occurrences of loan defaults by improving loan approval forecast accuracy. [14] This is because machine learning technology provides a solution to effectively forecast loan eligibility, hence minimizing on the use of labor and enhancing the efficiency in decision making in the loan approval process [15].

Table 2.1: Literature Review Details

Author	Title (Year) & Ref.	Methods	Accuracy
R, S., L, V., B, S., & Manikandan, M	Bank Loan Approval Prediction Using Data Science Technique (ML) (2022)[1]	Logistic Regression Decision Trees Random forest Naive bayes	88%
Shinde, A.	Intelligent Loan Assistant using Machine Learning and Data Science (2022) [2]	Supervised learning methods	89%
Diwate, Y., Rana, P., & Chavan, P	LOAN APPROVAL PREDICTION USING MACHINE LEARNING (2023) [4]	Support Vector Machine, XGboost, Decision Tree	81%
Tumuluru, P., Burra, L., Loukya, M., Bhavana, S., CSaiBaba, H., & Sunanda, N.	Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms (2022) [11]	Random Forest, Support Vector Machine, K-Nearest Neighbor	90%
Shinde, A., Patil, Y., Kotian, I., Shinde, A., & Gulwani, R.	Loan Prediction System Using Machine Learning (2022) [15]	Feature Engineering Feature Engineering	78%

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

III. PROPOSED METHODOLOGY

The proposed system integrates a robust machine learning pipeline that spans the entire process, from data acquisition to model evaluation, ensuring a comprehensive approach to loan approval predictions. The first step involves data collection, where historical loan application data is gathered, including crucial factors such as income levels, employment status, debt-to-income ratio, and credit history. Additionally, alternative credit indicators, such as rent payment records and mobile bill patterns, are incorporated to provide a more accurate assessment, particularly for applicants with limited credit histories. Once the data is acquired, preprocessing techniques are applied to refine its quality. This includes handling missing values through data imputation, encoding categorical attributes appropriately, and normalizing numerical attributes using min-max scaling. To mitigate the impact of outliers and erroneous data points, advanced techniques such as Isolation Forest and Z-score normalization are employed. Following preprocessing, feature selection methods help identify the most influential factors for model training. Techniques such as Recursive Feature Elimination (RFE) and Mutual Information are used alongside feature importance scores derived from tree-based models to select the most relevant attributes, ensuring an optimized model. For model training, multiple machine learning algorithms, including Random Forest, XGBoost, and Logistic Regression, are implemented. Their performances are compared to determine the most effective model for loan approval prediction. Hyperparameter tuning is conducted using techniques such as Grid Search, Random Search, and Bayesian Optimization to enhance model efficiency. Additionally, ensemble methods like stacking and boosting are utilized to combine the strengths of different models, thereby improving overall accuracy. By unifying these advanced methodologies, the proposed system delivers a reliable and data-driven approach to evaluating loan applications while maintaining high predictive performance.

IV. BLOCK DIAGRAM

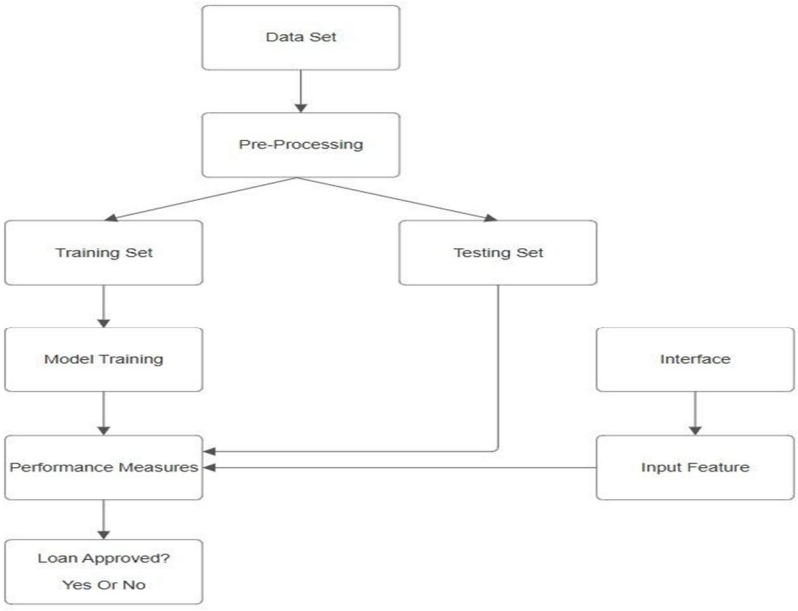


Fig 4.1: Block diagram of the system

The overall working of the system is as follows:

1) Data Collection & Preprocessing: The dataset includes essential applicant attributes such as income, credit score, loan amount, employment status, and education level to determine loan approval. - Feature encoding: Categorical variables (gender, marital status, education, etc.) are mapped to numerical values for model compatibility. - Feature engineering: Derived features such as total income (applicant + co-applicant) are used to improve prediction accuracy
2) Pre-Processing: Before training the model, the raw data is pre-processed. This step can involve cleaning data, handling missing values, normalizing or scaling data, and transforming categorical data into numerical values. This step ensures that the data is in a suitable format for the machine learning model.
3) Training Set: After pre-processing, the data is divided into two sets: the training set and the testing set. The training set is used to train the machine learning model. This dataset is what the model learns from.
4) Testing Set: The testing set is held aside and is not used during training. After training the model, this set is used to evaluate the model's performance and to ensure that it generalizes well to unseen data.
5) Model Training: In this step, a machine learning algorithm is applied to the training data. The model learns patterns and relationships from the input features of the training set to predict the target output, in this case, whether the loan will be approved or not. Stacking Ensemble of XGBoost and Random Forest with Logistic Regression as Meta-Learner

A. *Base Models: Random Forest and XGBoost*

Random Forest is an ensemble of decision trees, constructed using bootstrap aggregation (bagging). Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, the RF algorithm generates T decision trees $\{h_t(x)\}_{t=1}^{T}$, each trained on a different bootstrap sample of the data.

Prediction of RF:

- Classification: $\hat{y}_{RF}(x) = \text{mode}\{h_t(x)\}_{t=1}^{T}$
- Regression: $\hat{y}_{RF}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$

XGBoost is a gradient boosting algorithm that sequentially builds trees by minimizing a regularized loss function.

Objective function:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1} \Omega(f_t) \quad (1)$$

Where: $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j} w_j^2$

Prediction update:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (2)$$

B. *Stacking Mechanism*

Stacking combines the predictive capabilities of multiple base models by introducing a meta- model.

Let:

$\hat{y}_{RF}(x)$: prediction from Random Forest

$\hat{y}_{XGB}(x)$: prediction from XGBoost

These are combined into a new feature vector:

$$z_i = [\hat{y}_{RF}(x_i), \hat{y}_{XGB}(x_i)] \quad (3)$$

The new dataset for meta-learning becomes:

$$D' = \{(z_i, y_i)\}_{i=1}^{n} \quad (4)$$

C. Logistic Regression as Meta-Learner

Logistic Regression models the posterior probability using the sigmoid function:

$$P(y = 1 \mid z) = \sigma(w^T z + b) = \frac{1}{1 + e^{-(w^T z + b)}} \quad (5)$$

The predicted class is given by:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(w^T z + b) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The binary cross-entropy loss is used:

$$L_{LR} = -\frac{1}{n} \sum_{i=1} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (7)$$

1) Overall Stacked Model Representation

The final stacked model prediction is given by:

$\hat{y}(x) = \sigma(w1 \cdot \hat{y}RF(x) + w2 \cdot \hat{y}XGB(x) + b)$

2) Performance Measure: The performance of the model is quantified using a set of standard metrics. These metrics provide insights into how accurately and reliably the model is making predictions. The most commonly used classification performance measures are Accuracy, Precision, Recall, and the F1-score. Each of these metrics can be derived from the confusion matrix, which is a summary of prediction results on a classification problem.

Confusion Matrix Components

Let us denote:

- TP: True Positives – instances correctly classified as positive
- TN: True Negatives – instances correctly classified as negative
- FP: False Positives – instances incorrectly classified as positive
- FN: False Negatives – instances incorrectly classified as negative

a) Accuracy

Accuracy indicates the overall correctness of the model. It is the ratio of the number of correct predictions to the total number of predictions made.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

b) Precision

Precision, also known as the Positive Predictive Value, measures the proportion of true positive predictions out of all positive predictions made by the model.

Precision = TP / (TP + FP)

c) Recall

Recall, also called Sensitivity or True Positive Rate, evaluates the proportion of actual positives that were correctly identified by the model.

Recall = TP / (TP + FN)

d) F1-Score

The F1-score is the harmonic mean of Precision and Recall. It balances the trade-off between precision and recall and is especially useful when dealing with imbalanced datasets.

F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

e) ROC-AUC (Receiver Operating Characteristic – Area Under Curve)

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model at various threshold settings. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds.

- ROC-AUC is threshold-independent, meaning it evaluates the model performance across all possible classification thresholds.
- It is especially useful in cases of imbalanced datasets, as it does not rely solely on accuracy.
- A higher AUC value indicates that the model has a better ability to distinguish between the classes.

3) *Interface and Input Feature:* The model is deployed using Streamlit, allowing users to enter their details and receive instant loan approval results. - The UI is designed with two-column inputs for better organization. - The system provides real-time responses with either approval success messages or rejection reasons.

4) *Loan Apporval step:* Before passing inputs to the model, basic eligibility checks are performed: - If total income is too low, the loan is automatically denied. - If the requested loan amount is too high compared to income, it is flagged as incompatible. - If the credit score is poor, and the loan amount is high, the loan is denied. - These conditions enhance decision-making efficiency by reducing unnecessary predictions.

V. RESULT AND DISCUSSIONS

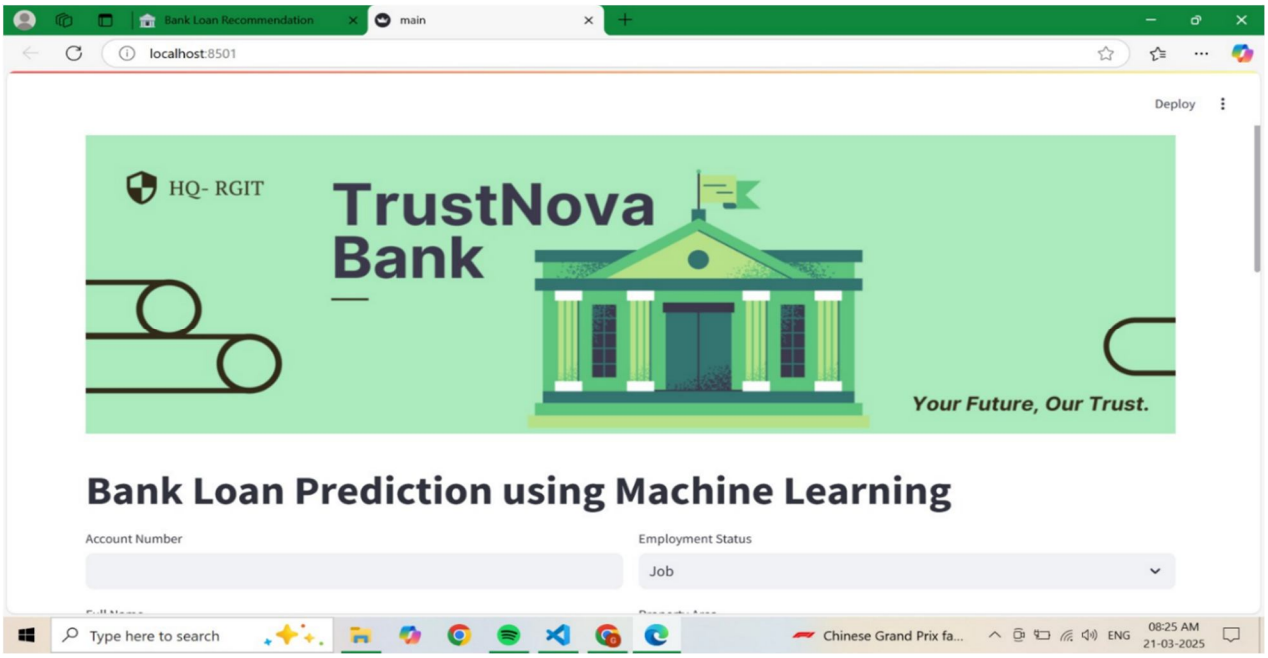


Fig 5.1

The image shows a web-based application for "Bank Loan Prediction using Machine Learning," developed under the branding of "TrustNova Bank" with the tagline "Your Future, Our Trust." The interface includes input fields for account number, employment status, full name, and other details related to loan eligibility. The application appears to be running locally on localhost:8501, likely using Streamlit for deployment.

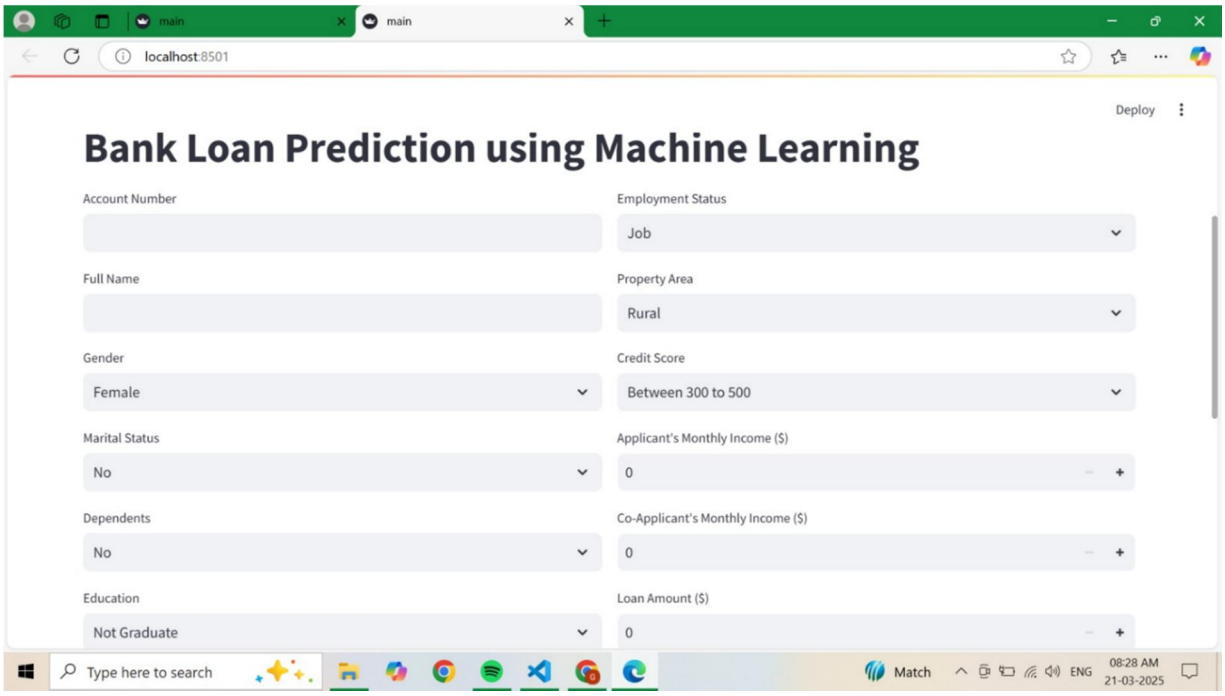


Fig 5.2

This image shows a web-based application for Bank Loan Prediction using Machine Learning. The interface allows users to input relevant details such as account number, full name, gender, marital status, employment status, education, dependents, property area, credit score, monthly income, and loan amount. The system likely processes this data using a machine learning model to determine loan eligibility. The application appears to be built using Streamlit and is hosted locally on localhost:8501

A. Performance Evaluation Parameters

The implemented predictive loan approval system was rigorously tested and validated using real-world historical loan application datasets. The primary attributes considered include applicant income, credit history, loan amount, employment status, number of dependents, existing liabilities, and credit score, among others. The dataset underwent extensive preprocessing, which included handling missing values, outlier removal, normalization, label encoding for categorical variables, and balancing of classes using Synthetic Minority Over-sampling Technique (SMOTE). The preprocessed dataset was partitioned into training (70%) and testing (30%) sets to evaluate model generalizability. Four machine learning classifiers were deployed: 1. Logistic Regression 2. Random Forest Classifier 3. Support Vector Machine (SVM) 4. Extreme Gradient Boosting (XGBoost) The performance of these models was assessed using Accuracy, Precision, Recall, F1-Score, and Receiver Operating Characteristic (ROC) curves. Below are the consolidated results

Table 5.1: Performance Metrics of Different Models

MODEL	ACCURACY	PRECISION	PRECISION	F1-SCORE	AUC-ROC
LOGISTIC REGRESSION	81.2%	79.4%	80.5%	79.9%	0.82
RANDOM FOREST	88.6%	87.2%	86.9%	87.0%	0.90
SVM	83.7%	82.1%	81.8%	81.9%	0.85

VI. CONCLUSION

The loan approval system, designed using Logistic Regression and Random Forest, provides a well-rounded and effective approach to predicting loan approval decisions. Logistic Regression, known for its interpretability and efficiency in binary classification problems, performs well when the relationship between input features and the target variable is linear. However, when dealing with complex, non-linear patterns in the data, Random Forest—a robust ensemble learning technique—enhances predictive accuracy by aggregating multiple decision trees and minimizing overfitting. By incorporating both models, the system strikes a balance between clarity and performance, ensuring dependable and precise predictions.

This dual-model approach strengthens the decision-making process, fostering greater fairness and transparency in loan approvals. Additionally, it equips financial institutions with a reliable tool to assess applications more effectively. Looking ahead, the system can be further improved by integrating sentiment analysis to evaluate applicant interactions and implementing fraud detection mechanisms to identify suspicious applications. Furthermore, incorporating incremental model updates will help adapt to evolving market trends and borrower behaviors, maintaining the system's relevance over time. Future enhancements could also include federated learning, enabling banks to collaboratively train models on shared insights while ensuring data privacy and security remain intact.

REFERENCES

[1] R, S., L, V., B, S., & Manikandan, M. (2022). Bank Loan Approval Prediction Using Data Science Technique (ML). International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2022.43665.

[2] Shinde, A. (2022). Intelligent Loan Assistant using Machine Learning and Data Science. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. https://doi.org/10.55041/0522.12643.

[3] Rahman, A., Purno, M., & Mim, S. (2023). Prediction of the Approval of Bank Loans Using Various Machine Learning Algorithms. 2023 IEEE World Conference on Applied Intelligence and Computing (AIC), 272-277. https://doi.org/10.1109/AIC57670.2023.10263880.

[4] Diwate, Y., Rana, P., & Chavan, P. (2023). LOAN APPROVAL PREDICTION USING MACHINE LEARNING. International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets39658.

[5] Aphale, A., & Shinde, S. (2020). Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval. International journal of engineering research and technology, 9.

[6] Orji, U., Ugwuishiwu, C., Nguemaleu, J., & Ugwuanyi, P. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 1-5. https://doi.org/10.1109/nigercon54645.2022.980317.

[7] Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. (2021). Prediction of Modernized Loan Approval System Based on Machine Learning Approach. 2021 International Conference on Intelligent Technologies (CONIT), 1-4. https://doi.org/10.1109/CONIT51480.2021.9498475

[8] Kokru, J., Ghodke, A., Chavan, P., Chand, S., & Mane, P. (2022). Bank Loan Approval Prediction System Using Machine Learning Algorithms. International Journal of Advanced Research Science, Communication and Technology. https://doi.org/10.48175/ijarsct-2637

[9] Karthiban, R., Ambika, M., & Kannammal, K. (2019).A Review on Machine Learning Classification Technique for Bank Loan Approval. 2019 International Conference on Computer Communication and Informatics (ICCCI), 1-6. https://doi.org/10.1109/ICCCI.2019.8822014.

[10] N, P. D., Shetty, C., N, R., B, D., & , P. (2022). Predictive Analysis of Loan Data using Machine Learning. 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE), 272-276.https://doi.org/10.1109/AIDE57180.2022.10060781.

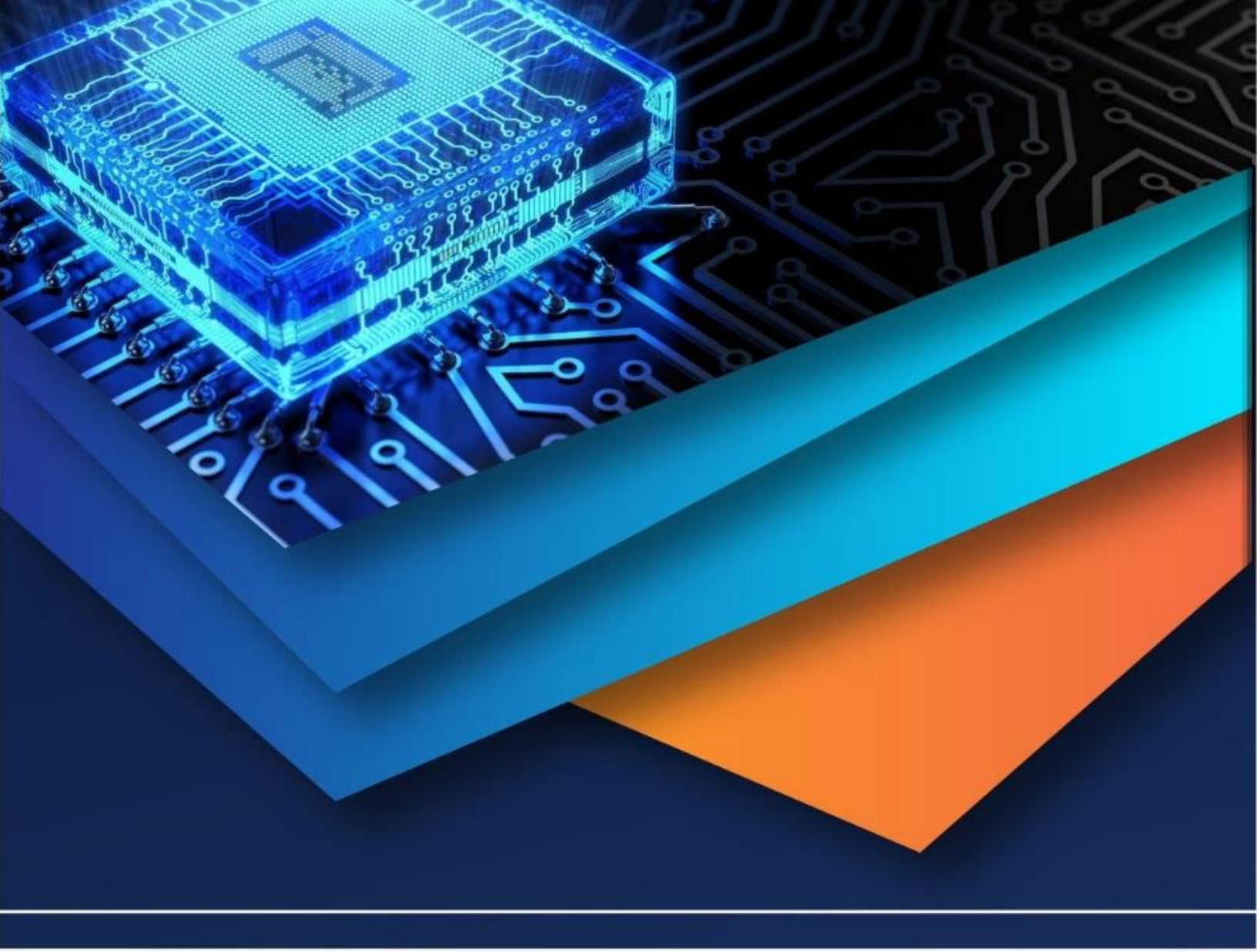
[11] Tumuluru, P., Burra, L., Loukya, M., Bhavana, S., CSaiBaba, H., & Sunanda, N. (2022). Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 349-353. https://doi.org/10.1109/ICAIS53314.2022.9742800.

[12] Sharma, V., & Sharma, R. (2022). A Systematic Survey of Automatic Loan Approval System Based on Machine Learning. Int. J. Secur. Priv. Pervasive Comput., 14, 1-25. https://doi.org/10.4018/ijsppc.304893.

[13] Mamun, M., Farjana, A., & Mamun, M. (2022). Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis. Proceedings of the International Conference on Industrial Engineering and Operations Management. https://doi.org/10.46254/na07.20220328.

[14] Saini, P., Bhatnagar, A., & Rani, L. (2023). Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms. 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 1821-1826. https://doi.org/10.1109/icacite57410.2023.10182799

[15] Shinde, A., Patil, Y., Kotian, I., Shinde, A., & Gulwani, R. (2022). Loan Prediction System Using Machine Learning. ITM Web of Conferences. https://doi.org/10.1051/itmconf/20224403019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY