International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 8, No. 5, (2025), pp. 302-311

Finance Loan Risk Assessment Using Machine Learning for Credit Eligibility Prediction and Model Optimization

Sigit Mulyanto[1], Dwika Lovitasari Yonia[2], Bambang Sutejo[3]*
[1, 3]Universitas Darwan Ali, Sampit, Kalimantan Tengah, Indonesia
[2]Institut Teknologi Sepuluh Nopember, Surabaya, Jawa Timur, Indonesia
Email: sigitmul@gmail.com[1], dwika.lovitasari.yonia@gmail.com[2], tejosampit@gmail.com[3]

Abstract

Finance loans play a crucial role in the global economy, supporting individuals and businesses in accessing capital for investment and financial stability. However, more than 60% of financial institutions struggle with identifying credit risks due to the limitations of traditional assessment models, which fail to capture complex borrower behavior. Additionally, 75% of financial firms have adopted AI-driven credit risk management, yet challenges remain in model validation and decision-making transparency. This study applies machine learning techniques, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), to enhance loan eligibility prediction. The dataset underwent preprocessing, including handling missing values, feature engineering, and data standardization. Model evaluation used accuracy, precision, recall, and F1-score. Logistic Regression achieved the highest F1-score at 88.6% and recall at 98.6%, while Random Forest recorded the highest precision at 82.3%. Feature importance analysis identified Credit History as the most influential factor, followed by Loan Amount and Total Income. While machine learning improves loan risk assessment, challenges remain in model interpretability. Future research should integrate explainable AI (XAI) and alternative credit scoring factors to enhance model transparency and robustness in real-world applications.

Keywords: *finance loan, credit risk assessment, machine learning, loan eligibility prediction, feature importance*

1. Introduction

Finance loans are vital to the global financial sector, yet over 60% of institutions struggle with credit risk assessment due to traditional models' limitations in capturing borrower behavior [1]. Loan eligibility assessment helps classify loan applicants based on four main criteria: payment ability, down payment, employment, and collateral, with the outcome being accepted, considered, or rejected [2]. Other research shows that the *Analytical Hierarchy Process* improves accuracy and efficiency in assessing credit eligibility based on economic conditions, character, capital, capacity, and collateral [3]. These models, often based on logistic regression, cannot handle large datasets or nonlinear relationships [4]. The rise of fintech and digital finance has increased the demand for automated credit evaluation, optimizing credit structures and reducing risk [5], [6]. Despite 75% of financial firms adopting AI-driven credit risk models, challenges persist in validation and transparency [7]. Ensemble techniques like XGBoost also enhance fraud detection and economic adaptability [8]. Expanding digital datasets allows feature engineering to uncover hidden credit patterns [9], though model generalizability and optimization remain challenges, requiring better hyperparameter tuning and feature selection [10]. Additionally, the book *Machine Learning in Data Science: Techniques and Cases* by Rudy C. Tarumingkeng

ISSN : 2580-7250

discusses how machine learning can be applied in the financial sector to improve credit assessment accuracy and detect transaction fraud [11].

This research aims to improve credit risk assessments by developing and comparing machine learning models, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, for loan approval prediction [12]. By evaluating model performance, the study identifies the most effective approach for optimizing loan approval accuracy while minimizing default risks, providing financial institutions with better risk management strategies and reduced default rates. In contrast to earlier studies that primarily emphasized accuracy, this research bridges gaps in comparative analysis, optimization methods for real-world datasets, and the interpretability of machine learning models within regulatory frameworks [13]. Additionally, it considers the adaptability of machine learning techniques to diverse borrower demographics, making it a crucial step toward bridging existing research gaps in credit assessment [5].

2. Research Methodology

The research methodology outlines the systematic approach used in developing a loan eligibility prediction model using machine learning. This section describes the step-by-step process, including data preparation, model selection, training, and evaluation. By following a structured approach, this study ensures the reliability, accuracy, and applicability of the predictive model in financial decision-making.

2.1. Research Workflow

The loan eligibility prediction process follows a structured methodology from data collection to model evaluation. Figure 1 outlines the workflow, starting with Data Preparation, where missing values are handled, categorical variables are encoded, and numerical attributes are normalized [14]. Model Development involves selecting machine learning models like Logistic Regression, Decision Tree, Random Forest, and SVM, with hyperparameter tuning using Grid Search or Random Search [15]. In Model Evaluation & Deployment, models are assessed using accuracy, precision, recall, and F1-score, identifying the best-performing model for financial applications [16] This approach enhances accuracy, efficiency, and transparency in loan approval automation.

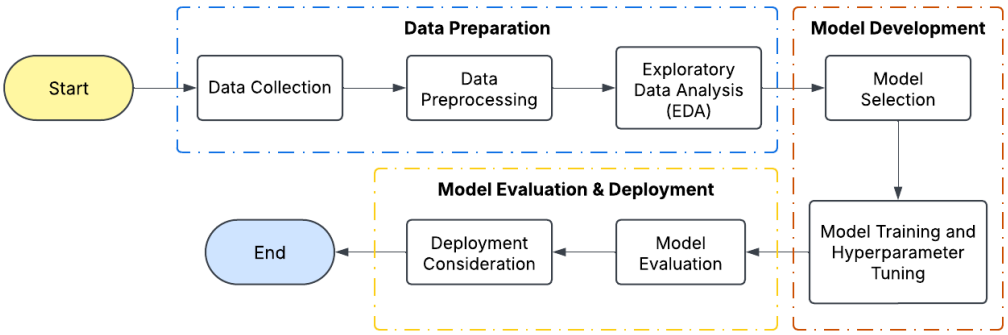

Figure 1. Research Workflow

2.2. Dataset Description

The "Finance Loan Approval Prediction Data" dataset, publicly available on Kaggle, contains key demographic and financial variables influencing loan eligibility and has not been widely used in previous studies. It includes applicant demographics, financial information, property details, and loan status, with data split into training and testing sets for model evaluation. Preprocessing techniques

such as handling missing values, encoding categorical variables, and standardizing numerical features were applied to improve model reliability, as summarized in Table 1.

Table 1. Dataset Attributes

No.	Attribute Name	Data Type	Description
1	Gender	Categorical	Applicant's gender (Male/Female)
2	Marital Status	Categorical	Marital status (Married/Single)
3	Dependents	Integer	Number of dependents (0, 1, 2, 3+)
4	ApplicantIncome	Numerical	Monthly income of the applicant
5	CoapplicantIncome	Numerical	Monthly income of the co-applicant
6	LoanAmount	Numerical	Requested loan amount
7	Loan_Amount_Term	Numerical	Loan term duration (in months)
8	Credit_History	Categorical	Applicant's credit history (1=favorable, 0=unfavorable)
9	Property_Area	Categorical	Location type (Urban, Semiurban, Rural)
10	Loan_Status	Categorical	Loan approval status (Y=Approved, N=Rejected)

2.3. Model Architecture

This study compares four machine learning models for loan eligibility prediction: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), which are widely used in financial risk assessment [17]. To enhance accuracy, hyperparameter tuning is performed using GridSearchCV or Random Search to find the best parameter combinations [18]. If class imbalance is detected, SMOTE (Synthetic Minority Over-sampling Technique) is applied to ensure fair loan approval predictions, as summarized in Table 2.

Table 2. Hyperparameters and Constraints

Model	Hyperparameters	Description & Constraints
Logistic Regression	C, solver	C (regularization) ∈ [0.01, 100], solver ∈ {'lbfgs', 'saga', 'newton-cg'}
Decision Tree Classifier	max_depth, criterion	max_depth ∈ [3, 50], criterion ∈ {'gini', 'entropy'}
Random Forest Classifier	n_estimators, max_depth	n_estimators (trees) ∈ [10, 500], max_depth ∈ [5, 100]
Support Vector Machine (SVM)	C, kernel	C (regularization) ∈ [0.01, 10], kernel ∈ {'linear', 'rbf'}

2.3.1. Logistic Regression

Logistic Regression is a common statistical model for binary classification, making it useful for predicting loan approvals based on applicant data. It calculates the probability of loan approval using a sigmoid function, where factors like income, credit history, and loan amount influence the decision [19]. In this model, $P(Y = 1)$ represents the probability of approval, while X_i are independent variables and β_i are coefficients that determine each predictor's impact, as shown in Equation (1).

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \quad (1)$$

2.3.2. Decision Tree Classifier

The Decision Tree Classifier predicts loan approvals by splitting the dataset into smaller subsets based on feature conditions [20]. It evaluates data using the Gini Index, which measures class impurity, selecting features that minimize impurity for

optimal splits. The process continues until the stopping criteria, such as maximum depth or minimum samples per leaf, are met, as shown in Equation (2).

$$Gini = 1 - \sum_{i=1}^{n} \left(p_i^2\right) \quad (2)$$

2.3.3. Random Forest Classifier

The Random Forest Classifier improves prediction accuracy by combining multiple Decision Trees in an ensemble learning approach [21]. Instead of relying on a single tree, it aggregates predictions from $\boldsymbol{N}$ independent trees, each trained on a randomly selected subset of data. The final classification is determined by averaging the outputs of all trees, reducing overfitting and improving generalization, as shown in Equation (3).

$$Prediction = \frac{1}{N}\sum_{i=1}^{N} Tree_i(X) \quad (3)$$

2.3.4. Support Vector Machine

The Support Vector Machine (SVM) classifies loan applications by finding an optimal hyperplane that maximizes the margin between approved and rejected loans. It introduces slack variables (ξ_i) to allow some misclassification, ensuring a balance between margin maximization and error minimization, where w is the weight vector, C is the regularization parameter, and ξ_i are the slack variables controlling misclassification penalties [22]. The formula used in this model is expressed in Equation (4).

$$\arg min \sum_{i=1}^{n} \left(\frac{1}{2}\omega^2 + C\sum \xi_i\right) \quad (4)$$

2.4. Model Training and Evaluation

The model training process involves data preprocessing, where missing values are handled, categorical variables are encoded, and numerical features are standardized. Machine learning models, including Logistic Regression, Decision Tree, Random Forest, and SVM, are trained using K-Fold cross-validation to ensure generalization. Hyperparameter optimization using GridSearchCV or Random Search fine-tunes essential parameters, including regularization strength (C), tree depth, and the number of estimators. Model evaluation is conducted using accuracy, precision, recall, and F1-score, selecting the best-performing model for real-world deployment [23].

3. Results and Discussion

This chapter presents the study's findings, analyzing the performance of machine learning models in predicting loan eligibility. The results are organized into data preparation, model development, and final assessment for deployment. A detailed discussion compares model performance, identifies key features, and examines challenges in automated loan approval systems.

3.1. Dataset Preparation

Data preparation ensures a clean dataset by handling missing values and conducting exploratory data analysis (EDA). Mode imputation was used for categorical variables, while mean imputation replaced missing numerical values to

maintain data consistency. Figures 2 and 3 illustrate numerical and categorical data visualizations, showing feature distributions, outliers, and frequency patterns for key variables.

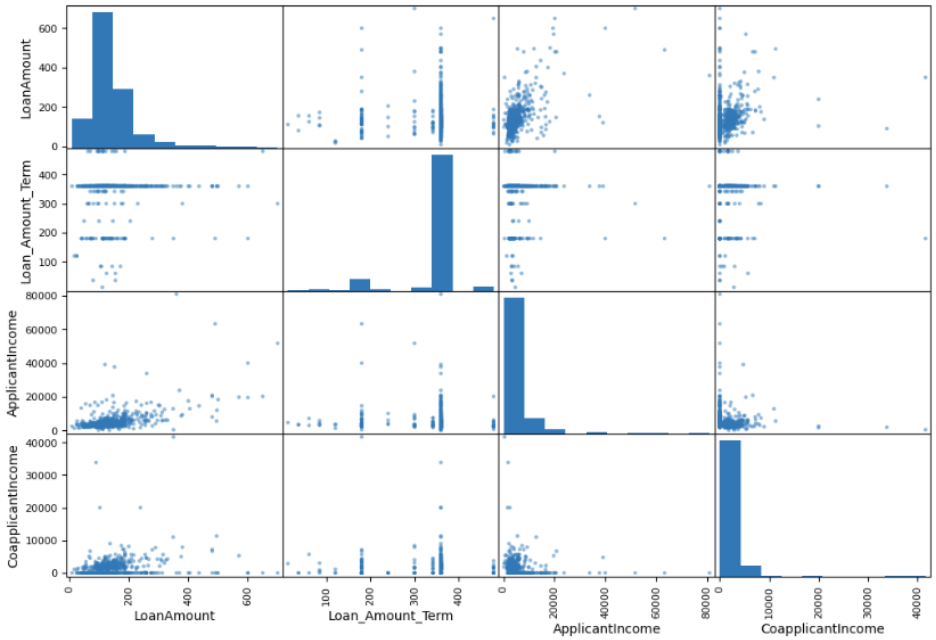

Figure 2. Numerical Data Visualization

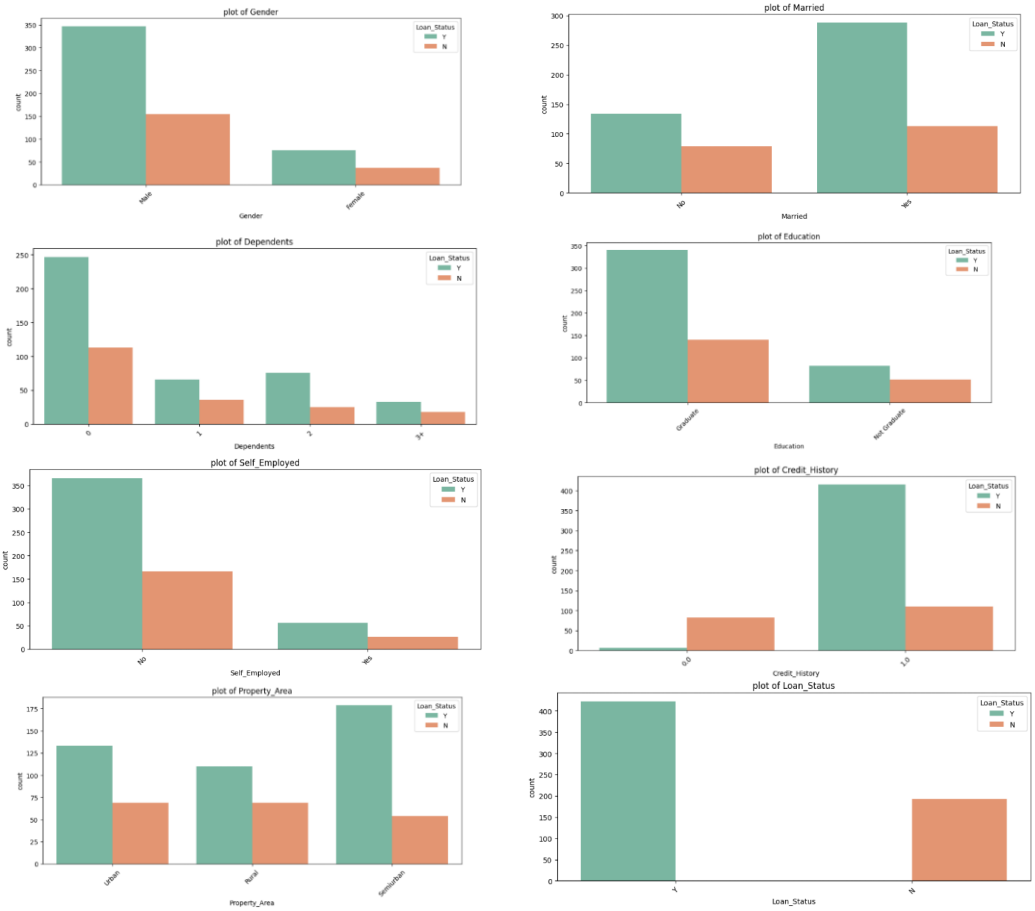

Figure 3. Categorical Data Visualization

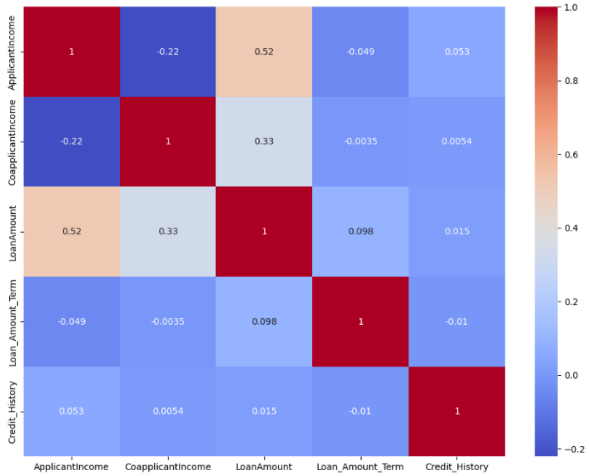

Figure 4. Correlation Analysis

Outlier detection using the Interquartile Range (IQR) method was applied to refine data quality, with extreme values either removed or capped. Correlation analysis, shown in Figure 4, revealed that ApplicantIncome and LoanAmount had a moderate correlation (0.52), while Credit_History remained a strong predictor despite weak correlations with numerical features. Feature engineering involved encoding categorical variables, standardizing numerical features, and creating new attributes like Total_Income and Loan-to-Income Ratio, before splitting the dataset into 80% training and 20% testing for model evaluation.

3.2. Model Development, Evaluation, and Deployment

Machine learning models for loan eligibility prediction were developed by selecting, training, optimizing, and evaluating four classifiers: Logistic Regression, Decision Tree, Random Forest, and SVM. Hyperparameter tuning using GridSearchCV and Random Search optimized key parameters, while K-Fold Cross-Validation ensured model robustness. The trained models were evaluated based on accuracy, precision, recall, and F1-score, with Table 3 summarizing the results, highlighting the best in bold and second-best performances in underline.

Table 3. Evaluation Results

Model	Metrix Evaluation			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Logistic Regression	**82,7%**	80,4%	**98,6%**	**88,6%**
Decision Tree	72,7%	81,6%	77,3%	79,4%
Random Forest	81,8%	**82,3%**	93,3%	87,4%
SVM	80,9%	78,7%	**98,6%**	87,5%

Logistic Regression achieved the highest F1-score of 88.6% and recall of 98.6%, making it the most effective model for identifying approved loans. Its high recall ensures minimal false negatives, improving loan approval predictions. However, Random Forest recorded the highest precision at 82.3% and a strong recall of 93.3%, making it highly reliable for minimizing false positives. Meanwhile, SVM also performed well, achieving a high recall of 98.6% and an F1-score of 87.5%, making it a competitive alternative. Overall, Logistic Regression is the best model due to its balanced precision and recall, but Random Forest is a strong choice for reducing incorrect approvals.

3.3. Feature Importance Analysis

Feature importance analysis, shown in Figure 5, identifies Credit History as the most influential factor in loan eligibility predictions. While Loan Amount and Total Income also contribute, their impact is significantly lower, with features like Property Area and Education playing minimal roles. These findings suggest that financial institutions prioritize Credit History in assessing loan risk, emphasizing the need for a strong credit record for successful loan applications.

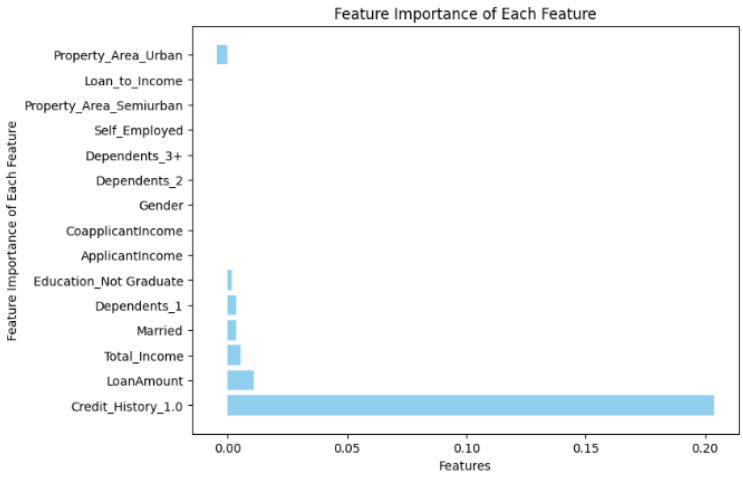

Figure 5. Feature Importance in Prediction

3.4. Baseline Model Performance

The baseline model serves as a benchmark for evaluating advanced machine learning models in loan eligibility prediction, with Logistic Regression selected due to its simplicity, interpretability, and efficiency in handling binary classification tasks. Trained using default parameters, it achieved an accuracy of 82.7%, precision of 80.4%, recall of 98.6%, and an F1-score of 88.6%, as shown in Table 3. The high recall indicates its effectiveness in identifying loan approvals, reducing false negatives, but its lower precision compared to Random Forest suggests a higher risk of false positives. Despite these limitations, Logistic Regression provides a strong reference point for assessing improvements made by more complex models. However, its inability to capture non-linear relationships and complex feature interactions highlights the need for more advanced models such as Random Forest and Support Vector Machine (SVM), which incorporate hyperparameter tuning, feature selection, and ensemble learning techniques to enhance predictive accuracy and reliability in financial decision-making systems.

4. Conclusion

This study investigated the application of machine learning models in predicting loan eligibility, evaluating Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) to determine the most effective approach. The findings confirm that machine learning enhances credit risk assessment, providing financial institutions with a data-driven alternative to traditional scoring methods. Logistic Regression emerged as the most balanced model, achieving the highest F1-score of 88.6% and recall of 98.6%, while Random Forest demonstrated superior precision at 82.3%, making it more reliable in minimizing false approvals. The analysis revealed that Credit History was the most influential factor in loan approval decisions, followed by Loan Amount and Total Income, highlighting the importance of financial behavior in credit assessment. A notable finding is that while models

performed well on historical data, their predictive power may be affected by economic fluctuations and regulatory changes, which were not directly considered in this study. Although machine learning proves effective, challenges remain in model interpretability and ethical concerns related to algorithmic decision-making. Future research should focus on integrating explainable AI (XAI) techniques to improve transparency and trust in automated credit evaluations while optimizing hyperparameter tuning and feature selection methods to enhance model adaptability in real-world financial applications.

References

[1] S. Mestiri and S. Mestiri, "Credit scoring using machine learning and deep Learning-Based models," *Data Science in Finance and Economics 2024 2:236*, vol. 4, no. 2, pp. 236–248, 2024, doi: 10.3934/DSFE.2024009.

[2] R. R. Yusran, "Criteria For Giving Car Loans to Consumers Using Sugeno's Fuzzy Concept," *IJISTECH (International Journal of Information System and Technology)*, vol. 6, no. 3, pp. 346–352, Oct. 2022, Accessed: Feb. 20, 2025. [Online]. Available: https://ijistech.org/ijistech/index.php/ijistech/article/view/248

[3] N. Gulo, E. Nurninawati, R. A. R. S, and D. P. Kristiadi, "Decision Support System for Submitting Credit using Analytical Hierarchy Process (AHP) Method Based on Android on Save and Loan Cooperatives Cubg Pasar Kemis Tangerang," *IJISTECH (International Journal of Information System and Technology)*, vol. 6, no. 4, pp. 441–448, Dec. 2022, Accessed: Feb. 20, 2025. [Online]. Available: https://ijistech.org/ijistech/index.php/ijistech/article/view/259

[4] A. Mukhanova *et al.*, "Forecasting creditworthiness in credit scoring using machine learning methods," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 5, pp. 5534–5542, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5534-5542.

[5] Z. Zhihui, "Impact of Digital Finance on Credit Structure and Risk-Taking in Commercial Banks: An Empirical Analysis," *International Journal of Education and Humanities*, vol. 4, no. 3, pp. 338–349, Jul. 2024, doi: 10.58557/(IJEH).V4I3.251.

[6] N. Metawa, R. Itani, S. Metawa, and A. Elgayar, "The impact of digitalization on credit risk: the mediating role of financial inclusion (National Bank of Egypt (NBE) case study)," *Economic Research-Ekonomska Istraživanja*, vol. 36, no. 2, Dec. 2023, doi: 10.1080/1331677X.2023.2178018.

[7] M. O. Kotb, "Credit Scoring Using Machine Learning Algorithms and Blockchain Technology," *1st International Conference of Intelligent Methods, Systems and Applications, IMSA 2023*, pp. 381–386, 2023, doi: 10.1109/IMSA58542.2023.10217411.

[8] N. Suhadolnik, J. Ueyama, and S. Da Silva, "Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach," *Journal of Risk and Financial Management 2023, Vol. 16, Page 496*, vol. 16, no. 12, p. 496, Nov. 2023, doi: 10.3390/JRFM16120496.

[9] T. Mokheleli and T. Museba, "Machine Learning Approach for Credit Score Predictions," *Journal of Information Systems and Informatics*, vol. 5, no. 2, pp. 497–517, May 2023, doi: 10.51519/JOURNALISI.V5I2.487.

[10] K. Yang, L. Liu, and Y. Wen, "The impact of Bayesian optimization on feature selection," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–11, Feb. 2024, doi: 10.1038/s41598-024-54515-w.

[11] O. : Prof, R. C. Tarumingkeng, and G. Besar, "Machine Learning dalam Data Science: Teknik dan Kasus".

[12] J. Quan and X. Sun, "Credit risk assessment using the factorization machine model with feature interactions," *Humanities and Social Sciences Communications 2024 11:1*, vol. 11, no. 1, pp. 1–10, Feb. 2024, doi: 10.1057/s41599-024-02700-7.

[13] S. Shi, R. Tse, W. Luo, S. D'Addona, and G. Pau, "Machine learning-driven credit risk: a systemic review," *Neural Comput Appl*, vol. 34, no. 17, pp. 14327–14339, Sep. 2022, doi: 10.1007/S00521-022-07472-2/TABLES/6.

[14] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu, and P. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development, NIGERCON 2022*, 2022, doi: 10.1109/NIGERCON54645.2022.9803172.

[15] T. Sarkar, M. Rakhra, V. Sharma, and A. Singh, "An Empirical Comparison of Machine Learning Techniques for Bank Loan Approval Prediction," *Proceedings of International Conference on Communication, Computer Sciences and Engineering, IC3SE 2024*, pp. 137–143, 2024, doi: 10.1109/IC3SE62002.2024.10593355.

[16] K. P, R. S, and J. Jaiswal, "Comparing Machine Learning Techniques for Loan Approval Prediction," Mar. 2024, doi: 10.4108/EAI.23-11-2023.2343174.

[17] G. Üniversitesi *et al.*, "A Comparative Study of Loan Approval Prediction Using Machine Learning Methods," *Gazi University Journal of Science Part C: Design and Technology*, vol. 12, no. 2, pp. 644–663, Jun. 2024, doi: 10.29109/GUJSC.1455978.

[18] T. Sarkar, M. Rakhra, V. Sharma, and A. Singh, "An Empirical Comparison of Machine Learning Techniques for Bank Loan Approval Prediction," *Proceedings of International Conference on Communication, Computer Sciences and Engineering, IC3SE 2024*, pp. 137–143, 2024, doi: 10.1109/IC3SE62002.2024.10593355.

[19] Y. Jiang, "Predicting Loan Default: A Comparative Analysis of Multiple Machine Learning Models," *Highlights in Science, Engineering and Technology*, vol. 85, pp. 169–175, Mar. 2024, doi: 10.54097/10DK2M95.

[20] P. S. Saini, A. Bhatnagar, and L. Rani, "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 2023*, pp. 1821–1826, 2023, doi: 10.1109/ICACITE57410.2023.10182799.

[21] M. Z. Hussain *et al.*, "Bank Loan Prediction System Using Machine Learning Models," *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024*, 2024, doi: 10.1109/I2CT61223.2024.10543786.

[22] J. Ren, Y. Wang, and X. Deng, "Slack-Factor-Based Fuzzy Support Vector Machine for Class Imbalance Problems," *ACM Trans Knowl Discov Data*, vol. 17, no. 6, Mar. 2023, doi: 10.1145/3579050.

[23] D. Lovitasari Yonia, A. Irham, D. Oranova Siahaan, and I. Mahfud, "Learning Models for Software Feature Extraction from Disaster Tweets: A Comparative Study," *ICECOS 2024 - 4th International Conference on Electrical Engineering and Computer Science, Proceeding*, pp. 83–88, 2024, doi: 10.1109/ICECOS63900.2024.10791279.