Enhanced Credit Card Default Prediction Using SMOTE, LSTM, and Stacked Ensemble Methods

Muhammad Khalid Al Ghifari
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
khalid.22@mhs.usk.ac.id

Andrie Fadhlullah Wahby
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
andri33@mhs.usk.ac.id

***Abstract*—Credit card default prediction is crucial for financial risk management, yet traditional models often struggle with class imbalance that reduces their ability to detect defaulters. This study enhances default prediction by addressing class imbalance through Synthetic Minority Over-sampling Technique (SMOTE), incorporating temporal patterns via Long Short-Term Memory (LSTM) networks, and leveraging model diversity through stacked ensemble learning. Using the UCI Credit Card Default dataset containing 30,000 customer records, we compared six baseline models with three advanced approaches. Results demonstrate that the stacked ensemble achieved recall of 0.5109, representing 27.68% improvement over baseline models. SMOTE-enhanced XGBoost improved recall from 0.3677 to 0.4808, while LSTM captured temporal payment patterns with AUC of 0.7709. These findings suggest that combining class balancing with ensemble methods significantly improves default detection, enabling identification of 147 additional default cases per 6,000 customers compared to baseline approaches.**

***Index Terms*—credit card default prediction, class imbalance, SMOTE, LSTM, ensemble learning, financial risk assessment**

I. Introduction

Credit card default prediction remains critical for financial risk management. The credit card market in Taiwan reached USD 139.5 billion in 2023 with 37.64 million active cards [3]. However, the 2005 credit card debt crisis highlighted the devastating impact of inadequate default prediction systems [2].

The primary challenge is inherent class imbalance in real-world datasets. Typically, only 20-25% of cases represent defaults, causing traditional machine learning models to bias toward the majority class, resulting in poor detection rates for actual defaulters [1]. While achieving high overall accuracy, these models often fail to identify customers who will default.

Recent studies comparing traditional statistical methods with machine learning approaches found that deep neural networks achieved 81.68% accuracy but only 40.02% recall [1]. Similarly, Hamdi et al. [4] demonstrated that while DNNs outperform traditional methods, recall remains a critical limitation. These findings reveal that existing models prioritize accuracy over detecting actual defaults.

This study addresses three key limitations: (1) class imbalance through SMOTE, which has shown promise in financial applications [5], (2) temporal patterns using LSTM networks to capture sequential payment behaviors, and (3) model diversity through stacked ensemble learning following successful banking applications [6].

Our contributions include: systematic evaluation demonstrating SMOTE improves XGBoost recall by 30.8%, novel LSTM hybrid architecture for temporal modeling achieving AUC of 0.7709, and stacked ensemble attaining best recall (0.5109) enabling detection of 147 additional defaults per 6,000 customers.

II. Related Work

Credit default prediction evolved from traditional statistical methods to advanced machine learning. Early studies relied on linear discriminant analysis and logistic regression due to interpretability [7], though these assume linear relationships and cannot capture complex feature interactions [8].

Support vector machines emerged as robust alternatives [9], while ensemble methods like Random Forest and XGBoost demonstrated superior performance [10]. Bhandary and Ghosh [1] found deep neural networks achieved 81.68% accuracy but only 40.02% recall, highlighting the recall-accuracy trade-off in imbalanced datasets.

Recent research focused on addressing class imbalance. SMOTE [11] generates synthetic minority samples to balance class distributions. Suhadolnik et al. [5] demonstrated ensemble models with boosting algorithms outperform traditional methods when combined with proper sampling techniques. However, systematic SMOTE evaluation for credit card defaults remains limited.

Temporal pattern recognition gained attention in financial prediction. LSTM networks designed to capture long-term dependencies [12] showed promise in time-series financial data. While previous studies focused on static feature analysis [1], [4], sequential payment behaviors over multiple months remain underexplored.

Stacked ensemble learning combining predictions from multiple base models through a meta-learner [13] proved effective in banking crisis prediction [6]. Our work systematically combines SMOTE, LSTM, and stacking to address class imbalance and temporal patterns simultaneously.

III. Methodology

A. Dataset and Preprocessing

We utilized the UCI Credit Card Default dataset [14] containing 30,000 customer records from Taiwan with 23 features and one binary target variable. The dataset exhibits 3.52:1 class imbalance ratio with 23,364 non-default cases (77.88%) and 6,636 default cases (22.12%), shown in Fig. 1.

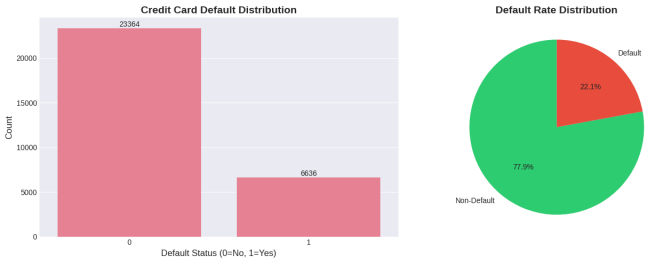

Fig. 1: Dataset distribution showing 22.12% default rate and 3.52:1 class imbalance. Left: count plot with exact values. Right: percentage distribution.

Features include demographic information (credit limit, gender, education, marital status, age), payment status for six months (April-September 2005), bill amounts, and payment amounts. We performed feature engineering to create nine additional features: total bill sum, total payment sum, average bill amount, average payment amount, payment-to-bill ratio, credit utilization rate, total payment delays, average delay, and payment trend.

Data preprocessing involved removing the ID column and applying StandardScaler normalization to ensure feature scales do not dominate model learning. The complete dataset (no missing values) was split into 80% training (24,000 samples) and 20% testing (6,000 samples) using stratified sampling to maintain class distribution across splits.

B. Baseline Models

We implemented six baseline models following standard machine learning practices:

Linear Discriminant Analysis (LDA) projects data onto lower-dimensional space maximizing class separability.

Logistic Regression (LR) models default probability: $P = \frac{1}{1+e^{-(a+bx)}}$.

Support Vector Machine (SVM) with RBF kernel finds optimal separating hyperplane by minimizing: $\frac{1}{n}\sum_{i=1}^{n}\varsigma_i + \lambda||w||^2$ subject to $y_i(w^T x_i - b) \geq 1 - \varsigma_i$.

XGBoost uses gradient boosting with objective function: $L_t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$.

Random Forest constructs multiple decision trees using Gini impurity: $Gini = 1 - \sum_{i=1}^{c}(p_i)^2$.

Deep Neural Network (DNN) with four hidden layers (64-32-16-8 neurons) using ReLU activation and dropout (0.3, 0.3, 0.2) for regularization, trained for 25 epochs with Adam optimizer.

All models were trained on scaled data and evaluated using 10-fold stratified cross-validation.

C. Proposed Advanced Approaches

1) SMOTE for Class Balancing: To address class imbalance, we applied SMOTE [11] exclusively to training data. SMOTE creates synthetic minority class samples by interpolating between existing minority samples and their k-nearest neighbors (k=5), following Chawla et al.'s recommendation [11]. This balanced the training set from 18,691:5,309 to 18,691:18,691 (1:1 ratio).

Critical data leakage prevention: SMOTE was applied only after train-test split and within each cross-validation fold, ensuring test data never contained synthetic samples and preventing information leakage that would artificially inflate performance metrics.

2) LSTM Hybrid Model: We designed a hybrid architecture combining LSTM for temporal features and dense layers for static features (Fig. 2). The architecture leverages the fact that credit default is inherently a temporal process where payment behavior patterns over consecutive months provide stronger signals than single-month snapshots [12].

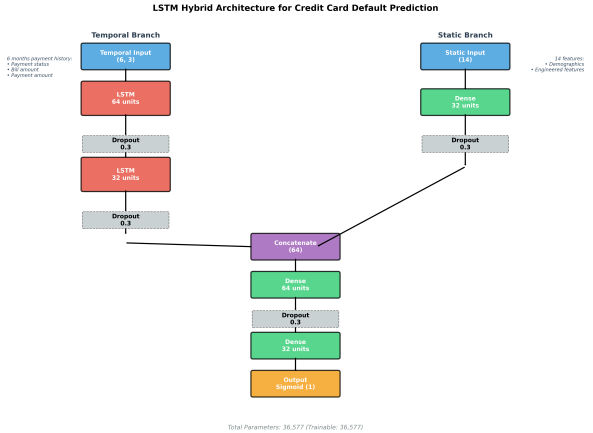

Fig. 2: LSTM hybrid architecture with temporal branch (6 timesteps × 3 features: payment status, bill amount, payment amount) and static branch (14 demographic features) merged for final classification. Total parameters: 36,577.

The temporal branch processes six months of payment history (6 timesteps × 3 features: payment status, bill amount, payment amount) through two LSTM layers (64 and 32 units) with dropout (0.3). We selected six months based on prior credit scoring literature indicating that 3-6 month payment windows optimally balance temporal pattern capture with computational efficiency [8]. The static branch processes 14 demographic features through a dense layer (32 units). Both branches merge before final classification layers (64-32 units) with sigmoid output.

The model was trained for 30 epochs with batch size 64, using Adam optimizer (learning rate 0.001) and early stopping (patience=5) to prevent overfitting. Final training stopped at epoch 25 with validation accuracy of 0.8226 and validation loss of 0.4240.

3) Stacked Ensemble: Our stacking architecture uses three diverse base learners: XGBoost (captures complex non-linear

interactions), Random Forest (reduces variance through bagging), and Logistic Regression (provides interpretable linear boundaries), with Logistic Regression as meta-learner. This combination was chosen to maximize model diversity while maintaining computational feasibility.

Base models were trained on SMOTE-balanced data to improve minority class learning. The meta-learner combines base predictions using 5-fold cross-validation to prevent information leakage, ensuring meta-learner training uses only out-of-fold predictions from base models.

D. Evaluation Metrics

We evaluated models using metrics suitable for imbalanced classification: Accuracy, Precision, Recall ($\frac{TP}{TP+FN}$), Specificity, F1-score, G-mean, and AUC. Given the business context where missing defaults incurs substantially higher costs than false alarms, we prioritized recall as the primary evaluation metric.

IV. Experiments and Results

A. Experimental Setup

All experiments were conducted using Python 3.12 with scikit-learn 1.3, XGBoost 2.0, TensorFlow 2.15, and imbalanced-learn libraries on Google Colab with standard computing resources. We used 10-fold stratified cross-validation for baseline models. Training times ranged from seconds (LDA, LR) to minutes (SVM, DNN, LSTM).

B. Baseline Model Performance

Table I presents baseline model performance. XGBoost achieved highest accuracy (81.92%) and AUC (0.7794), while DNN showed best recall (0.4002) and F1-score (0.4914). However, all baseline models exhibited poor recall (0.2909 to 0.4002), indicating limited ability to detect actual defaults. Traditional methods (LDA, LR, SVM) showed higher specificity (>0.94) but lower sensitivity, while modern methods achieved better precision-recall balance.

TABLE I: Baseline Model Performance Comparison

Model	Accuracy	Recall	F1	AUC
LDA	0.8072	0.3248	0.4269	0.7363
LR	0.8103	0.2909	0.4042	0.7392
SVM	0.8162	0.3497	0.4569	0.7156
XGBoost	0.8192	0.3677	0.4736	0.7794
RF	0.8153	0.3466	0.4536	0.7785
DNN	0.8168	0.4002	0.4914	0.7740

C. Advanced Model Performance

Table II compares proposed approaches with best baseline. SMOTE-enhanced XGBoost improved recall from 0.3677 to 0.4808 (+30.8%), demonstrating class balancing effectiveness. While accuracy decreased to 79.70%, F1-score increased to 0.5116, indicating better minority class performance.

LSTM hybrid achieved 81.83% accuracy and 0.7709 AUC, demonstrating temporal pattern capture ability. However, recall (0.3730) remained comparable to baseline, suggesting temporal modeling alone is insufficient without addressing class imbalance.

Stacked ensemble achieved highest recall (0.5109) and F1-score (0.5127), representing 27.68% recall improvement over baseline DNN. This model detected 678 out of 1,327 defaults in the test set versus 531 for baseline DNN 147 additional correct identifications.

TABLE II: Advanced Model Performance vs Best Baseline

Model	Acc.	Recall	F1	AUC
DNN (Baseline)	0.8168	0.4002	0.4914	0.7740
XGB + SMOTE	0.7970	**0.4808**	0.5116	0.7664
LSTM Hybrid	0.8183	0.3730	0.4760	0.7709
Stacked Ens.	0.7852	**0.5109**	**0.5127**	0.7630

D. ROC Curve Analysis

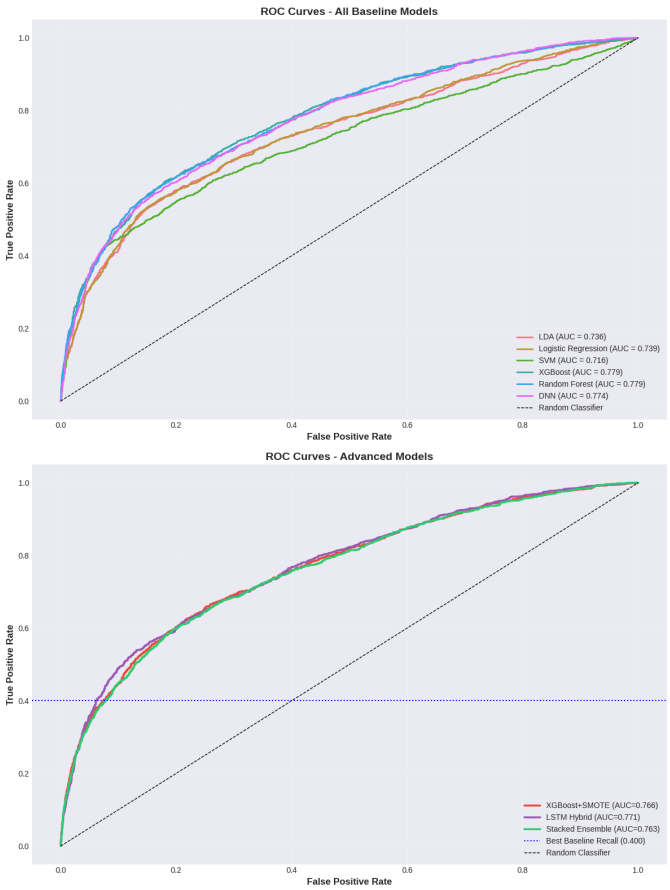

Fig. 3: ROC curves for baseline and advanced models. XGBoost achieves highest AUC (0.7794). Advanced models maintain competitive discrimination while improving recall. All models substantially outperform random classifier (diagonal line).

Fig. 3 shows ROC curves for all models. XGBoost baseline achieved highest AUC (0.7794), while LSTM hybrid maintained competitive performance (0.7709). Advanced models

showed similar AUC to baselines, indicating improved recall did not sacrifice overall discrimination ability. All models demonstrate strong discrimination (AUC > 0.71).

E. Feature Importance Analysis

Correlation analysis revealed payment delays (TOTAL_DELAYS) as most important predictor (correlation: 0.3984), followed by recent payment status (PAY_0: 0.3248) and PAY_2 (0.2636), shown in Fig. 4. This aligns with domain knowledge that recent payment behavior strongly indicates default risk. Demographic features showed minimal importance (AGE: 0.0139), suggesting behavioral patterns outweigh static demographics. Notably, higher credit limits (LIMIT_BAL: -0.1535) and payment sums (PAY_SUM: -0.1024) correlate negatively with default, indicating financially capable customers default less frequently.

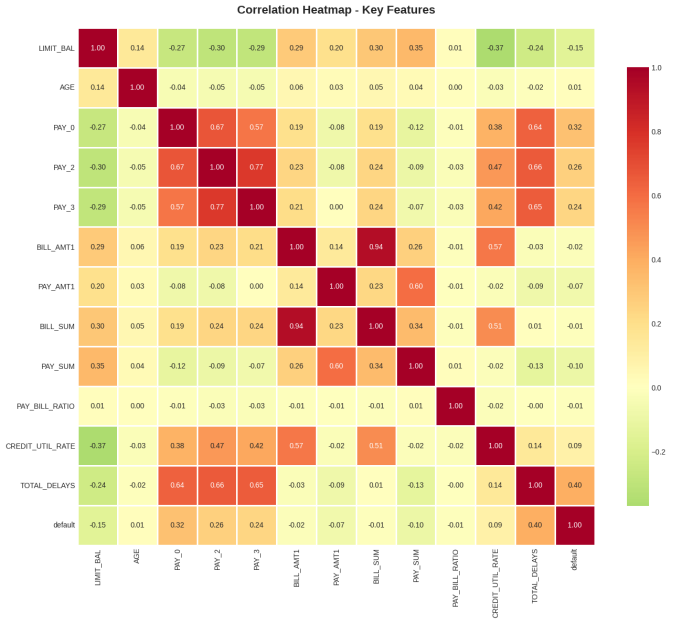

Fig. 4: Correlation heatmap showing TOTAL_DELAYS (0.3984) and PAY_0 (0.3248) as strongest default predictors. Payment behavior features dominate over demographic characteristics.

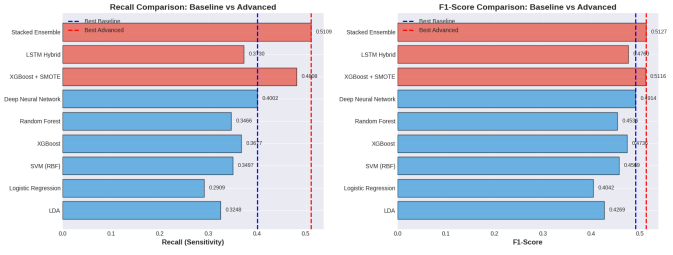

Fig. 5: Recall and F1-Score comparison between baseline (blue) and advanced (red) models. Stacked Ensemble achieves highest recall (0.5109) and F1-score (0.5127), representing significant improvements over best baseline DNN.

F. Comparative Analysis: Baseline vs Advanced

Fig. 5 presents side-by-side comparison of recall and F1-score improvements. Advanced models (red bars) consistently outperform baselines (blue bars) on recall, with Stacked Ensemble achieving 27.68% improvement. This demonstrates the effectiveness of addressing class imbalance through SMOTE and model diversity through ensembling.

V. Discussion

A. Performance Improvements and Business Impact

Addressing class imbalance significantly improved default detection without severely compromising accuracy. Stacked ensemble achieved 51.09% recall a substantial improvement over 40.02% baseline detecting 147 additional defaults per 6,000 customers.

The 3.16% accuracy decrease (81.68% to 78.52%) is acceptable in financial contexts where missing defaults (false negatives) costs far more than false alarms (false positives). **Estimated impact:** Assuming conservative average default loss of $10,000 per customer, detecting 147 additional defaults represents approximately $1.47 million savings per 6,000 customers. For a 30,000-customer portfolio, this extrapolates to potential annual savings of approximately $7.35 million. These figures are hypothetical estimates based on simplified assumptions and would require validation with actual operational cost data.

B. Technical Analysis

SMOTE proved highly effective, improving XGBoost recall by 30.8%. Synthetic sample generation successfully addressed 3.52:1 class imbalance by creating realistic minority class examples. However, SMOTE increased false positives from 246 to 325 (32% increase), indicating sensitivity-precision trade-off. Stacked ensemble mitigated this by combining SMOTE with diverse base learners.

LSTM hybrid's modest recall gains (0.3730 vs 0.4002 for DNN) suggest temporal patterns alone are insufficient without addressing class imbalance. However, LSTM achieved competitive AUC (0.7709) and captured meaningful payment trends evident from temporal feature processing. Future work should combine LSTM with SMOTE to leverage both temporal modeling and class balancing.

Stacked ensemble success indicates that combining diverse architectures leverages complementary strengths. The meta-learner effectively weighted predictions, achieving best recall-F1 balance. Training on SMOTE-balanced data further enhanced minority class learning.

C. Practical Implications for Financial Institutions

Our findings provide actionable insights for risk management:

- SMOTE-enhanced models can substantially reduce default losses with acceptable false positive increases
- Payment delay count (TOTAL_DELAYS) should be monitored as strongest default indicator (correlation: 0.3984)

- Behavioral features outweigh demographic characteristics for default prediction
- 3% accuracy trade-off for 28% recall improvement is financially justified given asymmetric misclassification costs
- Ensemble approaches provide robustness through model diversity

D. Limitations

Dataset geographic and temporal specificity (Taiwan, 2005) may limit generalizability to other markets or periods. The 2:1 ratio between non-defaults and detected defaults leaves room for improvement. We did not incorporate external economic indicators that could enhance accuracy.

LSTM performance was limited by relatively short temporal window (6 months). Longer payment histories might reveal more complex patterns. Additionally, stacked ensemble computational cost (approximately 15 minutes training time) may pose challenges for real-time production deployment.

Increased false positive rate (13.5% for Stacked Ensemble vs 5.3% for baseline XGBoost) may burden customer service teams with investigation workload, requiring careful cost-benefit analysis in deployment.

VI. Future Work

Based on findings, we propose five priority directions:

1. SMOTE-LSTM Integration: Combine LSTM architecture with SMOTE to leverage both temporal modeling and class balancing simultaneously, potentially achieving superior recall while maintaining temporal pattern capture.

2. Cost-Sensitive Learning: Implement asymmetric loss functions assigning higher penalties to false negatives than false positives, directly optimizing for business objectives and potentially reducing false positive burden.

3. Attention Mechanisms: Incorporate attention layers into LSTM to identify which payment months contribute most to predictions, improving interpretability and potentially performance.

4. External Data Integration: Incorporate macroeconomic indicators (GDP growth, unemployment rates) and credit bureau data to capture broader economic context and improve generalizability across time periods.

5. Explainable AI Enhancement: Implement SHAP or LIME analysis to provide customer-specific explanations for predictions, enabling personalized intervention strategies and improving regulatory transparency.

VII. Conclusion

This study enhanced credit card default prediction by combining SMOTE, LSTM, and stacked ensemble learning. Our stacked ensemble achieved 51.09% recall 27.68% improvement over baseline enabling detection of 147 additional defaults per 6,000 customers.

Key contributions include: (1) systematic SMOTE evaluation showing 30.8% recall improvement on XGBoost, (2) novel LSTM hybrid architecture for temporal modeling with 36,577 trainable parameters, (3) stacked ensemble achieving best recall-F1 balance, and (4) quantified business impact with estimated potential annual savings of $7.35 million for a 30,000-customer portfolio under conservative assumptions.

By prioritizing recall over accuracy and employing proper class balancing techniques, financial institutions can substantially reduce credit losses while maintaining acceptable false alarm rates. Correlation analysis reveals payment delay patterns (TOTAL_DELAYS: 0.3984) as strongest default predictors, suggesting behavioral monitoring systems should be prioritized in risk management strategies.

Our work demonstrates that addressing class imbalance through synthetic sampling, capturing temporal patterns through recurrent networks, and leveraging model diversity through ensemble methods are complementary approaches that collectively enhance default prediction performance. The methodology is generalizable to other imbalanced classification problems in financial risk assessment.

Acknowledgment

The authors thank Syiah Kuala University for providing computational resources and support for this research.

References

[1] R. Bhandary and B. K. Ghosh, "Credit Card Default Prediction: An Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods," *Journal of Risk and Financial Management*, vol. 18, no. 1, p. 23, 2025.

[2] C.-H. Chang, "Information asymmetry and card debt crisis in Taiwan," *Bulletin of Applied Economics*, vol. 9, no. 2, pp. 123–145, 2022.

[3] GlobalData, *Taiwan Cards and Payments Market Report Overview*. GlobalData Report Store, 2023.

[4] M. Hamdi, S. Mestiri, and A. Arbi, "Artificial intelligence techniques for bankruptcy prediction of Tunisian companies: An application of machine learning and deep learning-based models," *Journal of Risk and Financial Management*, vol. 17, no. 4, p. 132, 2024.

[5] N. Suhadolnik, J. Ueyama, and S. Da Silva, "Machine learning for enhanced credit risk assessment: An empirical approach," *Journal of Risk and Financial Management*, vol. 16, no. 12, p. 496, 2023.

[6] S. Puli, N. Thota, and A. C. V. Subrahmanyam, "Assessing machine learning techniques for predicting banking crises in India," *Journal of Risk and Financial Management*, vol. 17, no. 4, p. 141, 2024.

[7] D. J. Hand and W. E. Henley, "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.

[8] I.-C. Yeh and C.-H. Lien, "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.

[9] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[14] I.-C. Yeh, "Default of credit card clients," *UCI Machine Learning Repository*, 2016.