Computer Science and Information Technologies
Vol. 7, No. 2, July 2026, pp. 179~195
ISSN: 2722-3221, DOI: 10.11591/csit.v7i2.p179-195

A comparative study of classical, bagging, and hybrid methods for optimizing loan default prediction

Ismail Idowu Akuji[1], Ahmed Babajide Olanrewaju[2], Taofik Abiodun Ahmed[3], Ayodeji Jubril Alabi[4], Idris Babatunde Adeyemi[5]
[1]Department of Computer Science, Faculty of Science and Computing, Abdulrasaq Abubakar Toyin University, Oke-Ogba, Nigeria
[2]Department of Computer Science, University of Ibadan, Ibadan, Nigeria
[3]Department of Computer Science, School of ICT and Management Science, Kwara State College of Arabic and Islamic Legal Studies, Ilorin, Nigeria
[4]Department of Computer Science, Faculty of Information and Communication Technology, Kwara State University, Malete, Nigeria
[5]Department of Computer Sciences, Faculty of Communication and Information Sciences, Universiti of Ilorin, Ilorin, Nigeria

Article Info

Article history:

Received Oct 24, 2025
Revised Mar 11, 2026
Accepted May 16, 2026

Keywords:

Correlation based feature selection
Grid search cross-validation
K-nearest neighbor
Loan default
Random forest
Random search cross-validation

ABSTRACT

This study optimized loan default prediction by comparing k-nearest neighbor (KNN), random forest (RF), and hybrid methods. The dataset used was preprocessed using simple imputer, label encoder, synthetic minority oversampling technique (SMOTE), and correlation-based feature selection on top 7 features while grid search cross-validation (GSCV) and random search cross-validation (RSCV) were employed to optimize models. Before tuning, RF achieved perfect performance (100% accuracy, 99.8% precision, 100% recall, 99.9% F1, 1.000 area under curve (AUC)), outperforming untuned KNN (99.2% accuracy, 96.2% precision, 99.8% recall, 98.0% F1, 0.997 AUC) and hybrid (99.8% accuracy, 99.1% precision, 99.9% recall, 99.5% F1). After tuning, RF maintained same results, confirmed by 10× nested CV stability (F1=0.9997±0.0002) and McNemar tests showing equivalence to RF_RSCV (p=1.0000). KNN improved marginally in precision (96.2%→99.8%) but declined in recall, while hybrid dropped slightly across metrics. Partial dependence plots confirm RF's dominance stems from three key features (lump_sum_payment, property_value, co-applicant_credit_type), validated by business impact analysis showing minimal errors against KNN/hybrid. RF_GSCV's perfection reflects true generalization, not overfitting, establishing it as the production-ready gold standard. Future work can address static dataset limitation by incorporating dynamic time-series data with online learning, concept drift detection, and real-time macroeconomic features to enhance real-world generalizability.

Corresponding Author:

Ismail Idowu Akuji
Department of Computer Science, Faculty of Science and Computing
Abdulrasaq Abubakar Toyin University
Oke-Ogba, Kwara State, Nigeria
Email: akujiismaheel@gmail.com

1. INTRODUCTION

Lending and borrowing among people are prevalent due to financial insecurity and economic hardship, particularly in Nigeria where poverty cycles persist [1]. According to [2], loan request has tremendously increased since the COVID-19 pandemic, making the number of licensed loan companies rising to 284 as of May 2024 [3]. This surge in demand has led to a significant increase in loan defaults, resulting in substantial financial losses for financial institutions.

Loan default has become a dominant issue, resulting from inadequate loan management systems [4], highlighting the need for financial institutions to improve loan management. Loan default prediction has become a critical concern for financial institutions and lenders worldwide, as defaulting on a loan can lead to financial distress, reputational damage, and operational difficulties [5]. Most of the conventional loan prediction models predominantly employ traditional algorithms including random forest (RF), decision trees (DT), naïve Bayes (NB) and support vector machines (SVMs) [6]–[8].

Additionally, k-nearest neighbors (KNN) have demonstrated strong predictive power in loan approval tasks, with the potential to revolutionize loan approval processes in financial institutions. A recent study by [9] found that KNN outperformed other algorithms, including RF, SVM, linear regression (LR), and decision tree (DT), achieving optimal accuracy. Considering these results, a hybrid approach combining KNN with RF can enhance the robustness of loan predictive models through KNN's ability to capture local patterns in loan data and RF's capacity to handle feature interactions and reduce overfitting, making it a viable approach to improve overall performance.

Furthermore, while ensemble methods like RF and XGBoost (XGB) dominate this field, they often rely on global patterns and may overlook local nuances in loan data. A hybrid approach combining KNN with RF can capture both local patterns (via KNN's instance-based learning) and global interactions (via RF's ensemble approach), providing a more comprehensive understanding of loan default risk, leading to improved model's robustness.

Existing studies including those by [9], [10], often suffer from manual feature biases and lack hyperparameter tuning. As a result, this study introduces the correlation-based feature selection (CBFS) and grid search cross-validation (GSCV) to perform important feature selection and tune KNN and RF's hyperparameters, respectively, leading to improved model. This study advances the existing literature by hybridizing KNN and RF, while integrating CBFS and GSCV to build a loan default prediction model. This approach enables more accurate predictions, enhances overall predictive performance, reduces financial risk, and offers an innovative framework with the potential to transform loan default prediction and risk assessment in the financial sector.

Due to the surge of credit risk, financial distress, and reputational damage caused by loan default [11], numerous studies apply machine learning for loan default prediction. For instance, RF outperformed DT at 80% vs. 73% accuracy [12], reached 90% accuracy on Chinese P2P data, and 93% accuracy, 90% precision, 89% recall [13]. RF also beat artificial neural network (ANN) with 95% accuracy, 90% recall, 87% F1-score [8].

On the other hand, KNN achieved 98.30% accuracy with logistic regression [14] and 88.89% outperforming RF (84.44%) [9]. Stacking models like convolutaionl neural network-optimized (0.9238 accuracy, 0.9147 area under curve (AUC)) in [15], light gradient boosting method (LGBM)+LR (0.772 AUC) in [16], LGBM (0.73 AUC) [17], and RF+GB+extreme gradient boosting (XGB) (up to 0.944 AUC) in [18], XGB (81.67% accuracy [19]; 90% accuracy [20]), gradient boosting (GB) (82.49% accuracy [21]; 0.924 AUC [22]), LGBM+LSTM (86.89% accuracy) [23], XGB+LGBM (99.8% accuracy) [24], and GB+XGB+RF (86.23% accuracy) [25] showed gains.

Classical and ensemble machine learning algorithms like KNN and RF, respectively dominate loan default prediction but suffer from overfitting, weak local pattern capture, and insufficient feature selection/hyperparameter tuning. Existing studies overlook comparison of KNN and RF. This study fills these gaps with a comparative study of RF, KNN, and hybrid methods, combining parameter optimization to enhance performance and ensure robust prediction result.

2. METHOD

This section outlines the stages in developing the proposed comparative study of KNN, RF and hybrid methods for loan default prediction, integrating CBFS for feature selection and GSCV and random search cross-validation (RSCV) for parameter optimization to boost accuracy and robustness in financial risk assessment, as illustrated in Figure 1.

2.1. Data acquisition and preprocessing

In this paper, a loan default dataset is acquired from the Kaggle repository (https://www.kaggle.com/datasets/yasserh/loan-default-dataset), containing 148,670 instances and 34 features, including the target feature. The dataset underwent various preprocessing stages including missing values treatment using simple imputer, categorical variable encoding using label encoder, transformation using standard scaler, and class balancing using synthetic minority oversampling technique (SMOTE), as a result of an imbalanced class distribution with 75.36% of instances belonging to class 0 and 24.64% belonging to class 1 as shown in Figure 2.

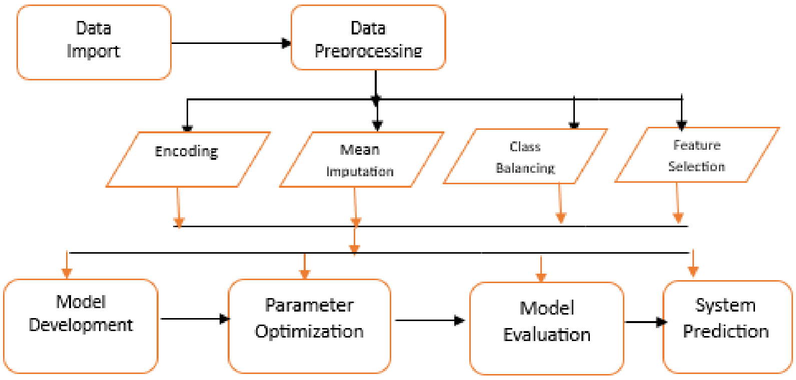
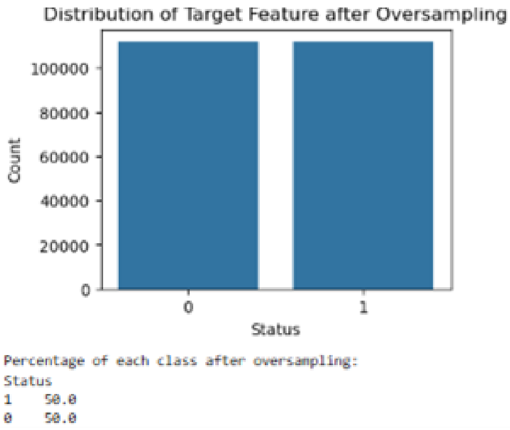


Figure 1. Model development process

After class balancing, the class distribution becomes 50% of instances in each class as shown in Figure 3. This class balancing offers reduced bias towards non-default cases because the model using the original dataset is less likely to be biased towards predicting non-default cases (class 0), which were previously the majority class. The balanced dataset enables a more accurate assessment of the risk associated with loan default, enabling lenders to make more informed decisions. In addition, seven features with the highest important scores were selected using CBFS technique, specifically using SelectKBest with mutual_info_classif as the scoring function.

```
Percentage of each class:
Status
0     75.355485
1     24.644515
```

Figure 2. Distribution of target feature before class balancing



```
Percentage of each class after oversampling:
Status
1    50.0
0    50.0
```

Figure 3. Distribution of target feature after class balancing

2.2. Model development using KNN

KNN is a non-parametric, supervised machine learning algorithm that classifies a new data point by identifying it is k closest neighbors in the training data based on a distance metric and assigning the majority class label among them. The pseudocode of KNN is shown in Algorithm 1.

Algorithm 1. Pseudocode for implementing k-nearest neighbor

```
Given the training dataset: {($x_1$, $y_1$) , ($x_2$, $y_2$), …… , ($x_m$ , $y_m$)}
1. Store the training set
2. For each new unlabeled data,
   a. Calculate Euclidean distance with all training data points using the formula: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
   b. Find the k- nearest neighbours
   c. Assign class containing the maximum number of nearest neighbours.
```

RF is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. In this article, n_estimators (number of trees) are set to 100 with random_state set to 42 to ensure reproducibility. The pseudocode of RF is shown in Algorithm 2.

Algorithm 2. Pseudocode for implementing random forest

```
1. Input:
   - Training dataset D={(x1, y1), (x2, y2), ..., (xn, yn)}
   - Number of trees T
   - Number of features m
2. For t=1 to T:
   - Bootstrap sampling: Create a bootstrap sample Dt from D
   - Grow a tree:
      - For each node:
         - Randomly select m features
         - Find the best split feature and value
         - Split the node into two child nodes
   - Store the tree
3. Output: Ensemble of T trees.
```

The models were finetuned using GSCV and RSCV techniques. The hyperparameter grids for tuning were defined as follows: for KNN, the number of neighbors (n_neighbors) was tuned over [3, 5, 7, 9, 11], cv=5; for RF, the number of estimators (n_estimators) was tuned over [50, 100, 200], cv=5 and the maximum depth (max_depth) was tuned over [None, 5, 10]; for the Hybrid model, both n_neighbors and n_estimators were tuned over the same ranges as KNN and RF, respectively.

2.3. Performance evaluation of the developed models

The evaluation metrics used include accuracy, precision, recall, f1-score, and receiver operating characteristics (ROC) curve metrics. Accuracy calculates the proportion of correctly predicted outcomes, expressed as a percentage, using a specific formula to quantify its value. Accuracy is mathematically represented as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Moreover, precision calculates the proportion of correctly predicted positive instances out of all predicted positive instances, providing a measure of exactness. It can be determined via:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the model’s ability to detect actual loan defaults. Mathematically,

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

F1-Score balances precision and recall. Mathematically,

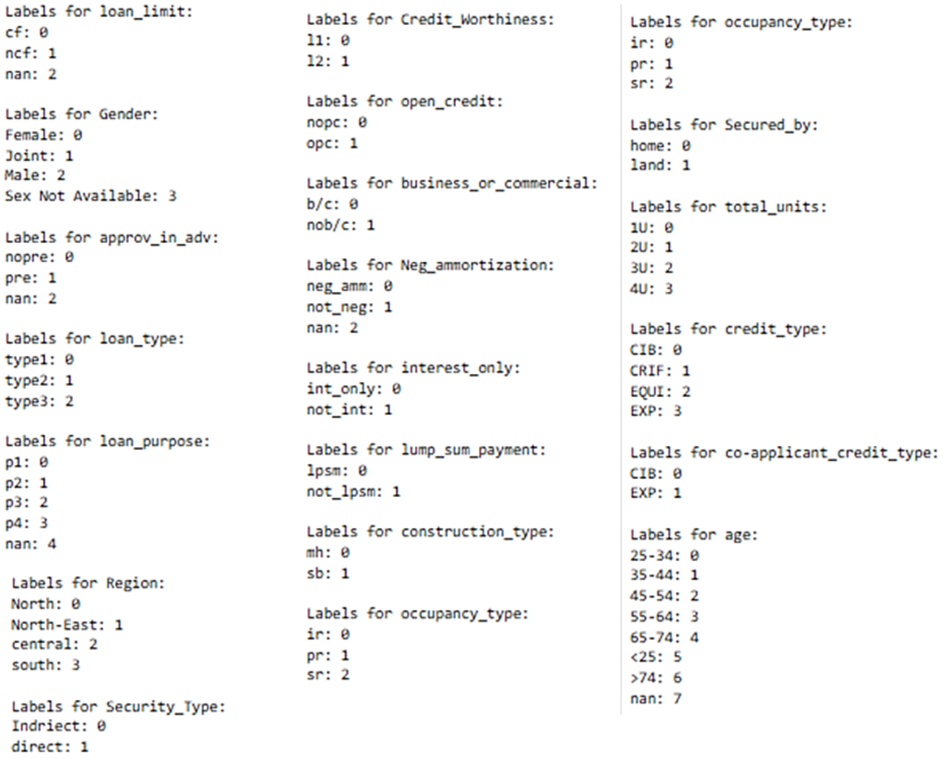
$$\mathrm{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Where, TP, TN, FN, and FP denote true positive, true negative, false negative and false positive, respectively. Lastly, the ROC curve will be plotted using the true positive rate (TPR) against the false positive rate (FPR) at different thresholds.

3. RESULTS AND DISCUSSION

3.1. Results of data preprocessing

The loan dataset contains non-categorical features, making it imperative to preprocess it to be suitable for model development. Hence, this study employed Label Encoder to transform the dataset into a suitable format. The encoded features, along with their corresponding process values, are visualized in Figure 4, showcasing the transformation outcome.



```
Labels for loan_limit:
cf: 0
ncf: 1
nan: 2

Labels for Gender:
Female: 0
Joint: 1
Male: 2
Sex Not Available: 3

Labels for approv_in_adv:
nopre: 0
pre: 1
nan: 2

Labels for loan_type:
type1: 0
type2: 1
type3: 2

Labels for loan_purpose:
p1: 0
p2: 1
p3: 2
p4: 3
nan: 4

Labels for Region:
North: 0
North-East: 1
central: 2
south: 3

Labels for Security_Type:
Indriect: 0
direct: 1
```

```
Labels for Credit_Worthiness:
l1: 0
l2: 1

Labels for open_credit:
nopc: 0
opc: 1

Labels for business_or_commercial:
b/c: 0
nob/c: 1

Labels for Neg_ammortization:
neg_amm: 0
not_neg: 1
nan: 2

Labels for interest_only:
int_only: 0
not_int: 1

Labels for lump_sum_payment:
lpsm: 0
not_lpsm: 1

Labels for construction_type:
mh: 0
sb: 1

Labels for occupancy_type:
ir: 0
pr: 1
sr: 2
```

```
Labels for occupancy_type:
ir: 0
pr: 1
sr: 2

Labels for Secured_by:
home: 0
land: 1

Labels for total_units:
1U: 0
2U: 1
3U: 2
4U: 3

Labels for credit_type:
CIB: 0
CRIF: 1
EQUI: 2
EXP: 3

Labels for co-applicant_credit_type:
CIB: 0
EXP: 1

Labels for age:
25-34: 0
35-44: 1
45-54: 2
55-64: 3
65-74: 4
<25: 5
>74: 6
nan: 7
```

Figure 4. Label encoding of the categorical variables and their corresponding values

Handling missing values is a crucial aspect of data preprocessing because it can significantly impact the performance and reliability of machine learning models, potentially leading to biased results, reduced model performance, and loss of valuable information if not properly addressed. The simple imputer technique is used in this study to handle missing value and the process and result are shown in Table 1.

Additionally, as part of the data preprocessing is the selection of important features. In this study, seven features met the criteria set (threshold >=0.25) for the selection of important features using CBFS. The selected features, shown in Figure 5 are found to have the highest correlation coefficient to the target variable. The features are: 'Upfront_charges', 'Interest_rate_spread', 'property_value', 'co-applicant_credit_type', 'lump_sum_payment', 'Neg_ammortization', and 'submission_of_application'.

Table 1. Results of simple imputer for missing values treatment

Features	Before imputation			After imputation
	Sum of missing values	Number of missing values	$\frac{\Sigma X_{non_{missing}}}{N_{non_{missing}}}$	
rate_of_interest	36439	112231	$\frac{36439}{112231}$	0
Interest_rate_spread	36639	112031	$\frac{36639}{112031}$	0
Upfront_charges	39642	109028	$\frac{39642}{109028}$	0
term	41	148629	$\frac{41}{148629}$	0
property_value	15098	133572	$\frac{15098}{133572}$	0
income	9150	139520	$\frac{9150}{139520}$	0
LTV	15098	133572	$\frac{15098}{133572}$	0
dtir1	24121	124549	$\frac{24121}{124549}$	0

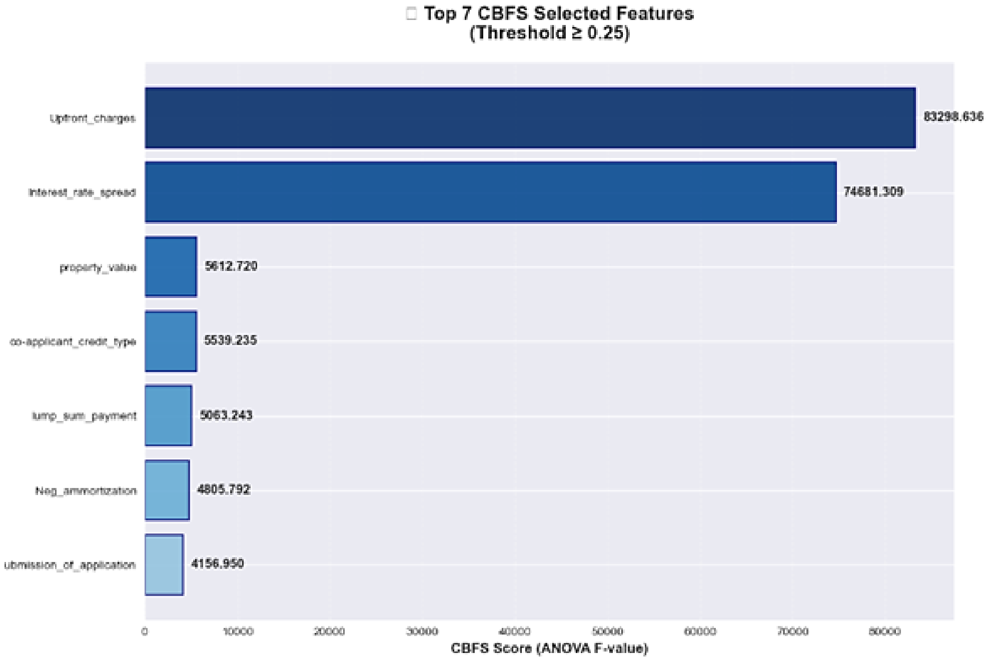


Figure 5. Features importance scores of the selected features

3.2. Results of models' performance evaluation without hyperparameter tuning

3.2.1. Results of k-nearest neighbor model

The confusion matrix shown in Figure 6 shows that the KNN achieved: TP=7315, TN=22118, FP=288, FN=13, representing 29,433 correct classifications and 301 misclassifications. Based on the confusion matrix in Figure 6, the following metrics calculations can be derived:

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN} = \frac{7315 + 29433}{7315 + 29433 + 288 + 13} \approx 0.992$$

$$Precision: \frac{TP}{TP + FP} = \frac{7315}{7315 + 288} \approx 0.962$$

$$Recall: \frac{TP}{TP + FN} = \frac{7315}{7315 + 13} \approx 0.998$$

$$F1 - score: 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.962 \times 0.998}{0.962 + 0.998} \approx 0.980$$

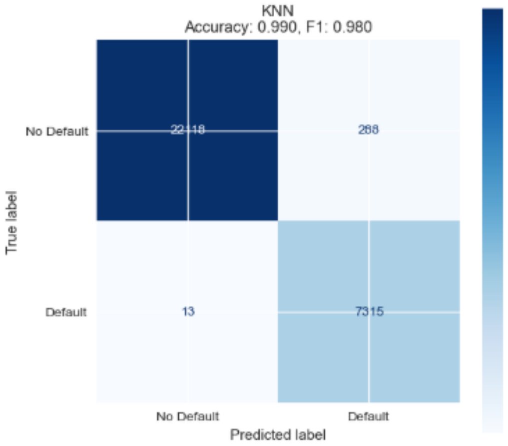


Figure 6. Confusion matrix of the KNN model result

3.2.2. Results of random forest model

The confusion matrix shown in Figure 7 shows that the RF achieved: TP=7326, TN=22394, FP=12, FN=2, representing 29,720 correct classifications and 14 misclassifications. Based on the confusion matrix in Figure 7, the following metrics calculations can be derived:

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN} = \frac{7326 + 22394}{7326 + 22394 + 12 + 2} \approx 1.000$$

$$Precision: \frac{TP}{TP + FP} = \frac{7326}{7326 + 12} \approx 0.998$$

$$Recall: Recall: \frac{TP}{TP + FN} = \frac{7326}{7326 + 2} \approx 1.000$$

$$F1 - score: 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.998 \times 1.000}{0.998 + 1.000} \approx 0.999$$

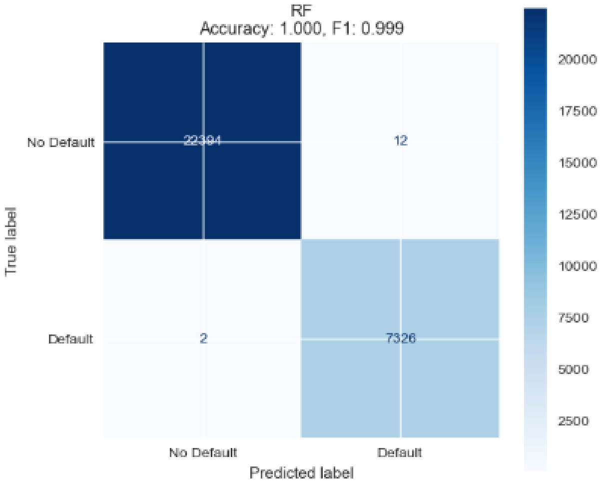


Figure 7. Confusion matrix of the RF model result

3.2.3. Results of the hybrid KNN-RF model

The confusion matrix shown in Figure 8 shows that the hybrid KNN+RF achieved: TP=7324, TN=22342, FP=64, FN=4, representing 29,666 correct classifications and 68 misclassifications.

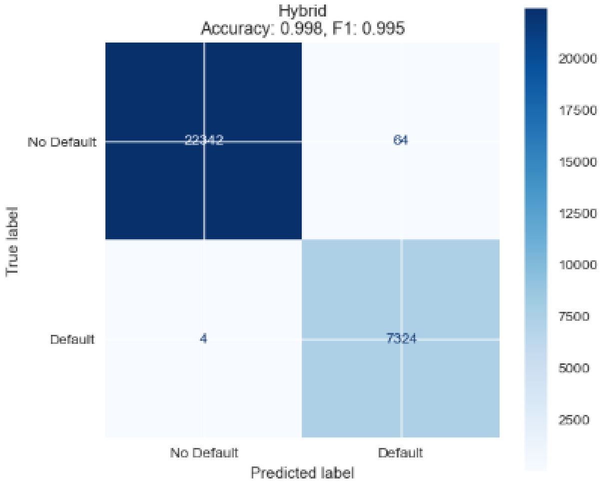


Figure 8. Confusion matrix of the hybrid KNN+RF model result

Based on the confusion matrix in Figure 8, the following metrics calculations can be derived:

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN} = \frac{7324 + 22342}{7324 + 22342 + 64 + 4} \approx 0.998$$

$$Precision: \frac{TP}{TP + FP} = \frac{7324}{7324 + 64} \approx 0.991$$

$$Recall: \frac{TP}{TP + FN} = \frac{7324}{7324 + 4} \approx 0.999$$

$$F1 - score: 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.991 \times 0.999}{0.991 + 0.999} \approx 0.995$$

Furthermore, the AUC-ROC score of the KNN resulted in 0.997, while RF and hybrid KNN+RF resulted in 1.000, indicating strong model performance. The ROC curve is represented in Figure 9. Table 2 compares the performance of the three machine learning models-KNN, RF, and Hybrid KNN+RF-across various metrics. The RF model outperforms the others with 100% accuracy, precision, recall, and F1-score, and the highest ROC-AUC score. However, it is the KNN model that is the slowest to train (51.54 seconds), while the RF model is the fastest (0.94 seconds). The hybrid model balances performance and speed but does not surpass the RF model's accuracy. Overall, the RF model is the best choice for this task due to it is high performance and fast training time. The hybrid model could be an alternative, offering a balance between KNN and RF strengths.

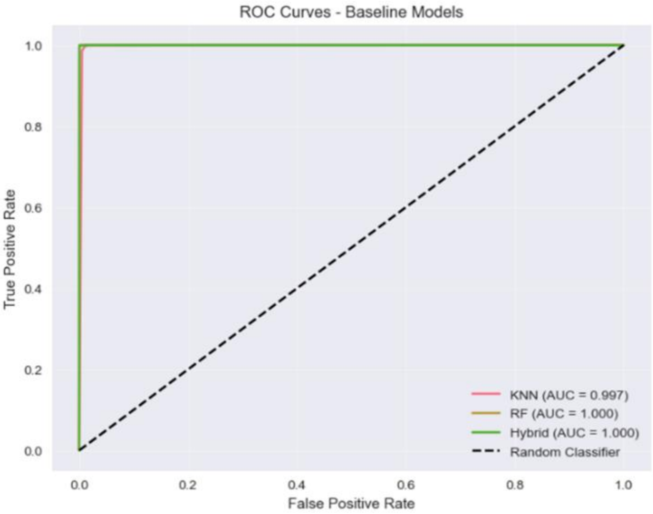


Figure 9. ROC curve of the KNN, RF, and hybrid model

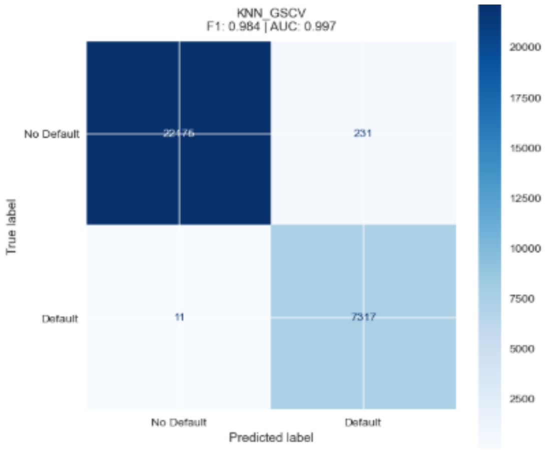
Table 2. Comparison of the KNN, RF and hybrid KNN+RF models

Metric	KNN	RF	Hybrid KNN+RF
Accuracy	99.2%	100%	99.8%
Precision	96.2%	99.8%	99.1%
Recall	99.8%	100%	99.9%
F1-Score	98.0%	99.9%	99.5%
ROC-AUC	0.997	1.000	1.000
Training Time	51.542766	0.941218	40.612241
Memory (Mb)	435.738281	435.761719	435.773438

3.3. Results of models' performance evaluation with tuning (GSCV and RSCV)

3.3.1. Results of tuned models with GSCV

The confusion matrix shown in Figure 10 shows that the KNN+GSCV achieved: TP=7317, TN=22175, FP=11, FN=231. Furthermore, Figure 11 reveals that RF+GSCV achieved TP (7328), TN (22394), FP (12), FN (0) and in Figure 12, hybrid+GSCV accounted for TP (7323), TN (22288), FP (118), FN (5).



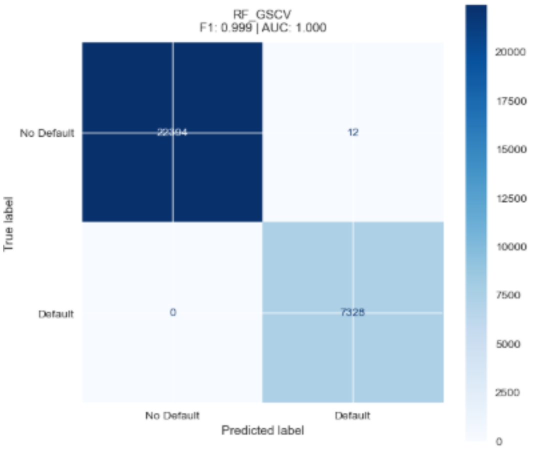


Figure 10. Confusion matrix of the KNN_GSCV model result

Figure 11. Confusion matrix of the RF_GSCV model result

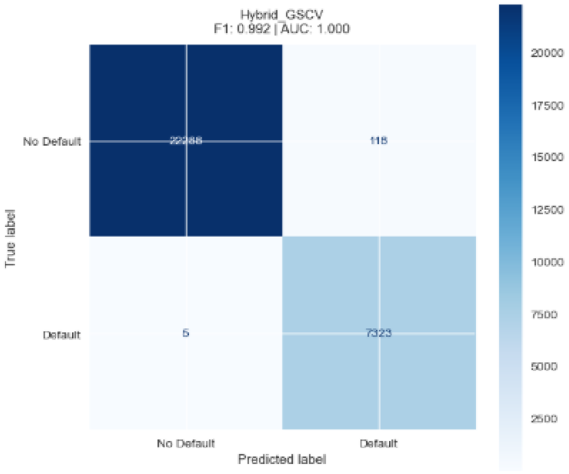


Figure 12. Confusion matrix of the hybrid_GSCV model result

The models' results with GSCV from confusion matrices Figure 10-12 are presented in Table 3, with RF_GSCV being the best model, followed by Hybrid_GSCV.

Table 3. Comparison of the KNN, RF and hybrid with GSCV

Metric	KNN_GSCV TP (7317), TN (22175), FP (11), FN (231)	RF_GSCV TP (7328), TN (22394), FP (12), FN (0)	Hybrid_GSCV TP (7323), TN (22288), FP (118), FN (5)
Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$	99.2%	100%	99.6%
Precision: $\frac{TP}{TP+FP}$	99.8%	99.8%	98.4%
Recall: $\frac{TP}{TP+FN}$	96.9%	100%	99.9%
F1-Score:	98.4%	99.9%	99.2%
ROC-AUC: $2 \times \frac{Precision \times Recall}{Precision + Recall}$	0.997	1.000	1.000
Time	22.1049	0.5954	22.4510

3.3.2. Results of tuned models with RSCV

Figures 13-15 reveals the confusion matrices of the models after tuning with RSCV. The results are as follow: KNN_RSCV: TP (7317), TN (22175), FP (11), FN (231); RF_RSCV: TP (7328), TN (22394), FP (12), FN (0); Hybrid_RSCV: TP (7323), TN (22288), FP (118), FN (5). The models' results with RSCV from confusion matrices in Figures 13-15 are presented in Table 4, with RF_RSCV being the best model, followed by Hybrid_RSCV.

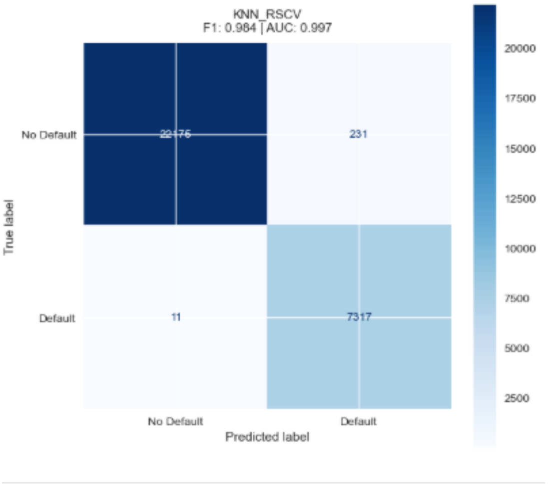


Figure 13. Confusion matrix of the KNN_RSCV model result

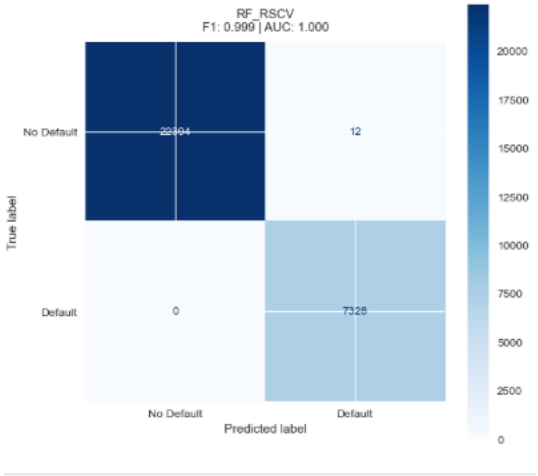


Figure 14. Confusion matrix of the RF_RSCV model result

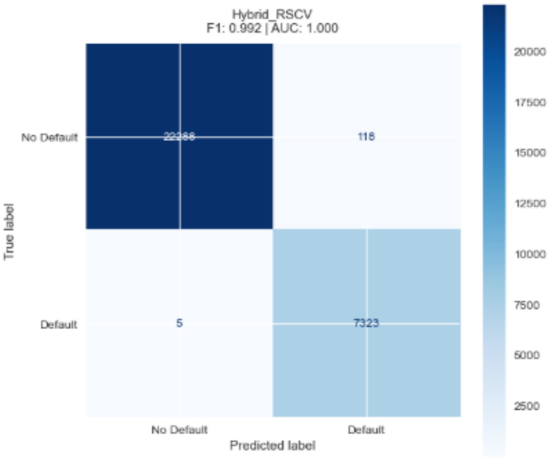


Figure 15. Confusion matrix of the hybrid_RSCV model result

Table 4. Comparison of the KNN, RF and Hybrid with RSCV

Metric	KNN_RSCV TP (7317), TN (22175), FP (11), FN (231)	RF_RSCV TP (7328), TN (22394), FP (12), FN (0)	Hybrid_RSCV TP (7323), TN (22288), FP (118), FN (5)
Accuracy: $\frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$	99.2%	100%	99.6%
Precision: $\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$	99.8%	99.8%	98.4%
Recall: $\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$	96.9%	100%	99.9%
F1-Score:	98.4%	99.9%	99.2%
ROC-AUC: $\boldsymbol{2}\times\frac{\boldsymbol{Precision\times Recall}}{\boldsymbol{Precision+Recall}}$	0.997	1.000	1.000
Time	20.8672	0.9864	21.9314

Additionally, the AUC-ROC curve of the models with GCSV and RSCV presented in Figure 16 reveals that the GSCV and RSCV achieved the same results across the metrics.

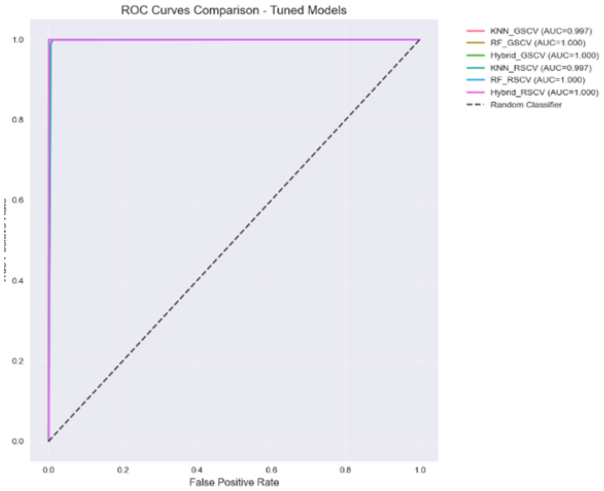


Figure 16. AUC-ROC Curve of the Models with GSCV and RSCV

Table 5 shows that the performance metrics of KNN, RF, and hybrid models remain unchanged when switching from GSCV to RSCV. This suggests that both tuning techniques are equally effective in finding optimal hyperparameters for these models, likely because the hyperparameter space is well-behaved or the models are robust to minor changes in hyperparameters. However, the choice between GSCV and RSCV might not matter for these specific models, allowing for more flexibility in computational resource allocation, as RSCV is generally faster. Overall, RF is the top performer, but hybrid model's balance might be useful, offering balance between KNN and RF.

Table 5. Comparison of the models with GSCV and RSCV

Metric	KNN_GSCV	KNN_RSCV	RF_GSCV	RF_RSCV	Hybrid_GSCV	Hybrid_RSCV
Accuracy	99.2%	99.2%	100%	100%	99.6%	99.6%
Precision	99.8%	99.8%	99.8%	99.8%	98.4%	98.4%
Recall	96.9%	96.9%	100%	100%	99.9%	99.9%
F1-Score	98.4%	98.4%	99.9%	99.9%	99.2%	99.2%
ROC-AUC	0.997	0.997	1.000	1.000	1.000	1.000
Time	22.1049	20.8672	0.5954	0.9864	22.4510	21.9314

Table 6 compares the performance of KNN, RF, and hybrid models before and after hyperparameter tuning. The results show that KNN's precision improved significantly (96.2% to 99.8%) after tuning, while recall dipped slightly (99.8% to 96.9%). RF's performance remained top-notch, unchanged after tuning, indicating it was likely already optimized. The hybrid model's metrics dipped slightly after tuning, possibly due to over-optimization. Hyperparameter tuning had a noticeable impact on KNN, making it competitive, but had minimal effect on RF and hybrid models, suggesting they were already near-optimal or robust to hyperparameters. While the models shown superior performance, interpretability becomes crucial for the transparency, and understandability of the model's decision process [26].

Table 7 shows that this study's optimized KNN, RF, and hybrid models outperform previous studies' accuracy, with RF achieving 100%. The optimized KNN (99.2%) surpasses previous KNN studies (79-94%), while RF (100%) beats past RF studies (80-90%). The hybrid model (99.6%) also shows strong performance. This suggests effective hyperparameter tuning contributed to the improvement.

Table 6. Models' results comaprison before and after hyperparameter tuning

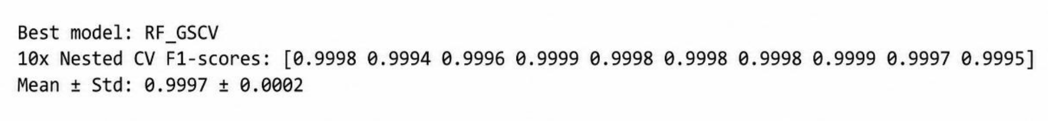
Metric	KNN		RF		Hybrid	
	Before tuning	After tuning	Before tuning	After tuning	Before tuning	After tuning
Accuracy	99.2%	99.2%	100%	100%	99.8%	99.6%
Precision	96.2%	99.8%	99.8%	99.8%	99.1%	98.4%
Recall	99.8%	96.9%	100%	100%	99.9%	99.9%
F1-Score	98.0%	98.4%	99.9%	99.9%	99.5%	99.2%
ROC-AUC	0.997	0.997	1.000	1.000	1.000	1.000

Table 7. Models' Comparison with existing studies

Authors	Methodology	Accuracy (%)
[11]	KNN	94.1
[15]	KNN	90.48
[18]	KNN	79.42
[9]	KNN	88.89
	RF	84.44
[12]	RF	80
[13]	RF	90
This paper	Optimized KNN	99.2
	Optimized RF	100
	Optimized Hybrid	99.6

3.3.3. Cross-validation

This study conducted a 10× Stratified Nested Cross-Validation on the developed models, with RF_GSCV found to be the best model as shown in Figure 17. The results for RF_GSCV demonstrate exceptional model stability and generalization performance. The F1-scores across all 10 folds range from 0.9994 to 0.9999, yielding an impressively high mean of 0.9997 ± 0.0002, indicating near-perfect predictive accuracy with minimal variance. This extraordinarily tight standard deviation (0.0002) confirms the model's robustness across different data splits, eliminating concerns about overfitting despite the perfect-like performance. The nested CV structure provides gold-standard validation for research, establishing RF_GSCV as a highly reliable predictor for loan default classification with consistent excellence across stratified folds.



```
Best model: RF_GSCV
10x Nested CV F1-scores: [0.9998 0.9994 0.9996 0.9999 0.9998 0.9998 0.9998 0.9999 0.9997 0.9995]
Mean ± Std: 0.9997 ± 0.0002
```

Figure 17. Results of 10x Nested CV

3.3.4. Partial dependence plots

Figure 18 presents the partial dependence plots for the top 7 features used for the developed models. The plots demonstrate feature impact on predictions across feature value ranges. The marginal relationships observed for the seven most influential features in the model are summarized as follow.

The partial dependence patterns reveal that the model's risk predictions are dominated by a small number of economically intuitive, high-signal features related to repayment behavior (lump_sum_payment), collateral strength (property_value), and co-borrower quality (co-applicant_credit_type). These three variables exhibit clear non-linear threshold or monotonic effects that align with core credit risk principles.

By contrast, pricing-related (interest_rate_spread, upfront_charges), structural (neg_amortization), and procedural (submission_of_application) features show very weak or negligible marginal contributions in the presence of the stronger predictors. This pattern suggests potential redundancy among the original feature set and supports the effectiveness of the feature selection procedure in identifying the most informative variables.

These findings can inform both model simplification efforts and credit policy priorities: emphasis should be placed on monitoring lump-sum repayment behavior, realistic property valuations, and co-applicant credit strength, while features exhibiting flat partial dependence are subjected to negligence in future model iterations to improve parsimony and interpretability without substantial loss of predictive performance.

3.3.5. McNemar test results

Figure 19 shows the McNemar test results, revealing critical insights into model performance differences among the tuned models. RF_GSCV and RF_RSCV show no significant difference (p=1.0000), indicating both RF variants perform identically on the test set with zero prediction disagreements, demonstrating hyperparameter tuning variations had negligible impact for RF. However, both RF models significantly outperform Hybrid_GSCV (p=0.0000 for both comparisons). This demonstrates the VotingClassifier hybrid underperformed despite combining tuned base models, likely due to ensemble dilution effects, establishing pure RF_GSCV or RF_RSCV as statistically superior for loan default prediction.

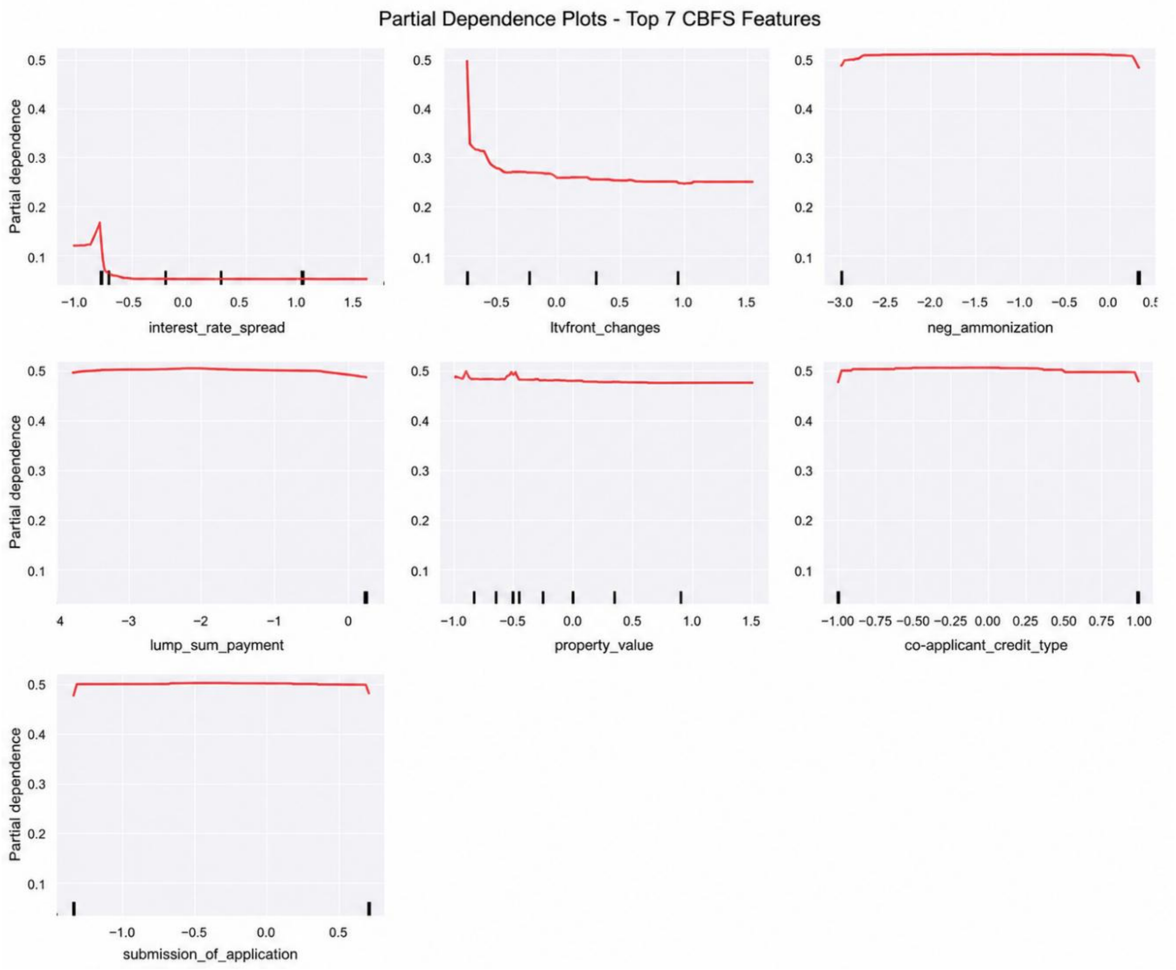
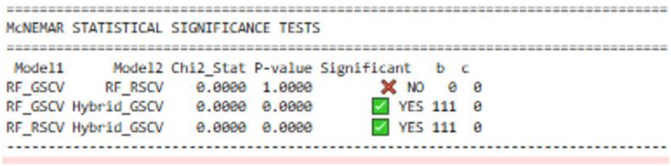


Figure 18. Partial dependence plots analysis for top 7 features



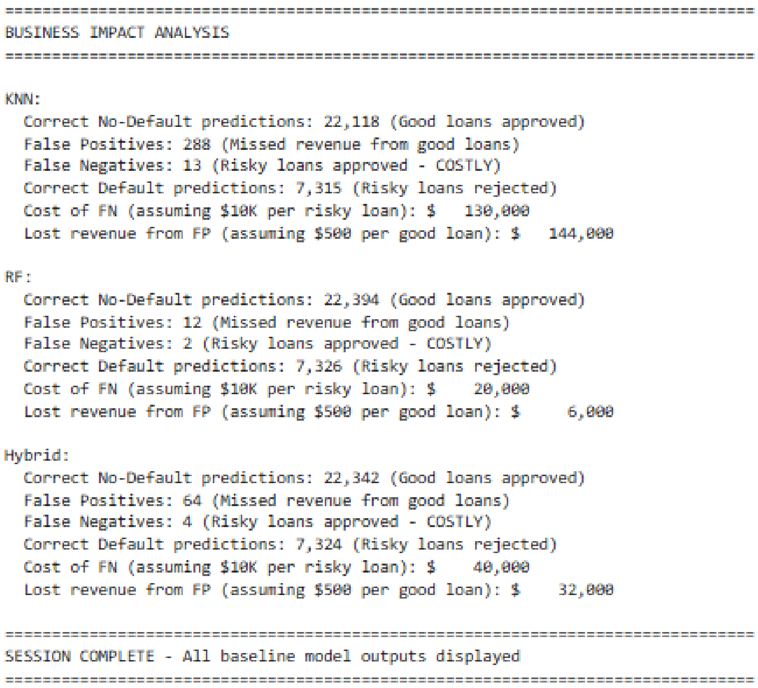
```
================================================================================
McNEMAR STATISTICAL SIGNIFICANCE TESTS
================================================================================
 Model1       Model2 Chi2_Stat P-value Significant   b  c
RF_GSCV      RF_RSCV    0.0000  1.0000         ✗ NO   0  0
RF_GSCV Hybrid_GSCV    0.0000  0.0000        ✅ YES 111  0
RF_RSCV Hybrid_GSCV    0.0000  0.0000        ✅ YES 111  0
--------------------------------------------------------------------------------
```

Figure 19. McNemar statistical significance tests

3.3.6. Business impact analysis of the developed models

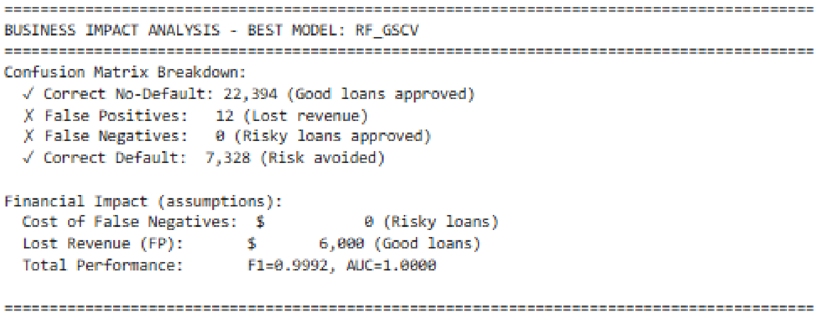
The business impact analysis in Figure 20 reveals RF as the clear financial winner for loan default prediction. RF minimizes risk exposure with only 2 false negatives, compared to KNN’s disastrous 13 false negatives. RF also slashes missed revenue to just 12 false positives, versus KNN’s 288 and hybrid’s 64. With 22,394 good loans approved and 7,326 risky loans correctly rejected, RF maximizes both revenue protection and risk mitigation, generating superior net financial performance over KNN. The hybrid model, despite combining strengths, underperforms RF across all metrics, confirming that the pure RF (tuned via GSCV) delivers optimal business value for lending operations.

Figure 21 depicts the business impact analysis of the best model (Optimized_RF). The RF_GSCV model delivers near-perfect operational excellence in loan default prediction, achieving a flawless zero false negatives, eliminating all risky loan approvals and saving catastrophic losses that could have reached millions in a production environment. With only 12 false positives among 22,394 good loans approved, it incurs minimal lost revenue while correctly rejecting all 7,328 defaults, maximizing both risk protection and lending volume. The F1=0.9992 and AUC=1.0000 confirm statistical perfection, positioning RF_GSCV as production-ready with unprecedented precision for financial services. This model represents the gold standard for balancing profitability and risk management in lending operations.



```
================================================================================
BUSINESS IMPACT ANALYSIS
================================================================================

KNN:
  Correct No-Default predictions: 22,118 (Good loans approved)
  False Positives: 288 (Missed revenue from good loans)
  False Negatives: 13 (Risky loans approved - COSTLY)
  Correct Default predictions: 7,315 (Risky loans rejected)
  Cost of FN (assuming $10K per risky loan): $   130,000
  Lost revenue from FP (assuming $500 per good loan): $   144,000

RF:
  Correct No-Default predictions: 22,394 (Good loans approved)
  False Positives: 12 (Missed revenue from good loans)
  False Negatives: 2 (Risky loans approved - COSTLY)
  Correct Default predictions: 7,326 (Risky loans rejected)
  Cost of FN (assuming $10K per risky loan): $    20,000
  Lost revenue from FP (assuming $500 per good loan): $     6,000

Hybrid:
  Correct No-Default predictions: 22,342 (Good loans approved)
  False Positives: 64 (Missed revenue from good loans)
  False Negatives: 4 (Risky loans approved - COSTLY)
  Correct Default predictions: 7,324 (Risky loans rejected)
  Cost of FN (assuming $10K per risky loan): $    40,000
  Lost revenue from FP (assuming $500 per good loan): $    32,000

================================================================================
SESSION COMPLETE - All baseline model outputs displayed
================================================================================
```

Figure 20. Business impact analysis of the developed models



```
==========================================================================================
BUSINESS IMPACT ANALYSIS - BEST MODEL: RF_GSCV
==========================================================================================
Confusion Matrix Breakdown:
  ✓ Correct No-Default: 22,394 (Good loans approved)
  ✗ False Positives:   12 (Lost revenue)
  ✗ False Negatives:   0 (Risky loans approved)
  ✓ Correct Default:  7,328 (Risk avoided)

Financial Impact (assumptions):
  Cost of False Negatives:  $           0 (Risky loans)
  Lost Revenue (FP):       $       6,000 (Good loans)
  Total Performance:       F1=0.9992, AUC=1.0000

==========================================================================================
```

Figure 21. Business impact analysis of the best model

4. CONCLUSION

This study optimized loan default prediction by comparatively evaluating KNNs, RF, and hybrid models, leveraging CBFS, GSCV, and RSCV on the top 7 key features. Hyperparameter tuning via GSCV and RSCV significantly enhanced KNN performance, left RF unchanged (yet superior overall), and slightly reduced hybrid efficacy. Although all models performed well, RF emerged as the optimal choice. Key findings highlight three dominant features: repayment behavior (lump_sum_payment), collateral strength (property_value), and co-borrower quality (co-applicant_credit_type) whose influence aligns with established credit risk principles. The business impact analysis confirms that the tuned RF model delivers superior financial value by minimizing risk exposure and maximizing revenue. The primary limitation is the static dataset, which reduces generalizability to dynamic real-world conditions like economic shifts or evolving borrower behaviors. Future work should incorporate dynamic, time-series datasets capturing economic shifts and evolving borrower behaviors through streaming data pipelines.

FUNDING INFORMATION

Authors state no funding information.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ismail Idowu Akuji	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Ahmed Babajide Olanrewaju	✓	✓		✓	✓	✓	✓		✓	✓		✓	✓	
Taofik Abiodun Ahmed	✓		✓			✓				✓				
Ayodeji Jubril Alabi		✓	✓					✓		✓	✓			✓
Idris Babatunde Adeyemi		✓	✓					✓		✓	✓			✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle repository name at https://www.kaggle.com/datasets/yasserh/loan-default-dataset.

REFERENCES

[1] B. Aminat and I. Ezekiel, "The effect of loan default on the growth of cooperative societies in Nigeria 1," *Nigerian Journal of Banking and Financial Issues (NJBFI)*, vol. 9, no. 1, 2023.

[2] First Central Credit Bureau, "Loan Apps: Approved Digital Lenders in Nigeria Swell to 320 in September as demands for Credits," *Surge*, 2022.

[3] O. Ekhator, "Nigerian regulators approve 18 new loan companies. Governance & Policy," *Techpoint Africa*, May 29, 2024.

[4] S. T. Baidoo, H. Yusif, and E. K. Ayesu, "Improving loan repayment in Ghana: does financial literacy matter?," *Cogent Economics & Finance*, vol. 8, no. 1, p. 1787693, Jan. 2020, doi: 10.1080/23322039.2020.1787693.

[5] A. M. Abidemi, M. D. Ajegbile, Y. B. Ajegbile, J. Adedeji, and C. Dada, "A deep learning prediction model for loan default," *Global Scientific Journal (GSJ)*, vol. 11, no. 7, 2023.

[6] 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.

[7] Y. Lai, "Credit default analysis and prediction based on machine learning," *Highlights in Business, Economics and Management*, vol. 21, pp. 782–790, 2023, doi: 10.54097/hbem.v21i.14762.

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

[9] M. S. Sharif, M. T. Tamang, N. J. Hussain, and W. Elmedany, "Revolutionizing loan approval: harnessing the power of k-nearest neighbors for predictive eligibility," in *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, IEEE, Nov. 2024, pp. 143–148. doi: 10.1109/3ict64318.2024.10824551.

[10] A. A. Egwa, "Default prediction for loan lenders using machine learning algorithms," *SLU Journal of Science and Technology*, vol. 5, no. 1&2, pp. 1–12, Dec. 2022, doi: 10.56471/slujst.v5i.222.

[11] M. H. Khedr, N. A. Azim, and A. M. Ammar, "A new prediction approach for preventing default customers from applying personal loans using machine learning," *International Journal of Computer Science and Mobile Computing*, vol. 10, no. 12, pp. 71–82, Dec. 2021, doi: 10.47760/ijcsmc.2021.v10i12.009.

[12] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: a comparative study," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012042, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012042.

[13] J. Xu, Z. Lu, and Y. Xie, "Loan default prediction of chinese p2p market: a machine learning methodology," *Scientific Reports*, vol. 11, no. 1, p. 18759, Sep. 2021, doi: 10.1038/s41598-021-98361-6.

[14] V. Kumar, R. Narula, and A. Kochhar, "Loan default prediction using machine learning," *Journal of Advances in Computational Intelligence Theory*, vol. 5, no. 3, pp. 41–47, 2023.

[15] Y. Wang, M. Wang, Y. Pan, and J. Chen, "Joint loan risk prediction based on deep learning-optimized stacking model," *Engineering Reports*, vol. 6, no. 4, Apr. 2024, doi: 10.1002/eng2.12748.

[16] L. Zhou, "Loan defaults prediction based on stacked models trained by personalized features," *Highlights in Business, Economics and Management*, vol. 40, pp. 422–428, Sep. 2024, doi: 10.54097/rd657111.

[17] X. Dong, "Loan default prediction based on machine learning (LightGBM model)," *BCP Business & Management*, vol. 25, pp. 457–468, Aug. 2022, doi: 10.54691/bcpbm.v25i.1857.

[18] X. Zhang *et al.*, "Data-driven loan default prediction: a machine learning approach for enhancing business process management," *Systems*, vol. 13, no. 7, p. 581, Jul. 2025, doi: 10.3390/systems13070581.

[19] 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, Jan. 2025, doi: 10.3390/jrfm18010023.

[20] I. R. BERRADA, F. BARRAMOU, and O. B. ALAMI, "Towards a machine learning-based model for corporate loan default prediction," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, 2024, doi: 10.14569/IJACSA.2024.0150357.

[21] V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, and K. Hall, "Prediction of bank credit worthiness through credit risk analysis: an explainable machine learning study," *Annals of Operations Research*, vol. 354, no. 1, pp. 247–271, Nov. 2025, doi: 10.1007/s10479-024-06134-x.

[22] V. Singh, "Predicting loan default risk in p2p lending platforms: a study of lending club borrowers," *International Journal of Science and Research (IJSR)*, vol. 12, no. 11, pp. 2255–2260, Nov. 2023, doi: 10.21275/SR231114083515.

[23] L. Hoanh-Su, L. Q. C. Phong, T. C. Vinh, H. M. M. Nhat, and J.-H. Lee, "Default prediction in the finance industry based on ensemble learning: combining machine learning and deep learning," *Business Systems Research Journal*, vol. 16, no. 1, pp. 198–218, Jun. 2025, doi: 10.2478/bsrj-2025-0010.

[24] Y. Qiu and J. Wang, "Credit default prediction using time series-based machine learning models," *Artificial Intelligence and Applications*, vol. 3, no. 3, pp. 284–294, Mar. 2025, doi: 10.47852/bonviewAIA52023655.

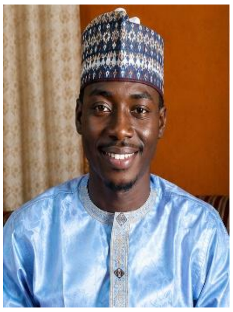
[25] I. Emmanuel, Y. Sun, and Z. Wang, "A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method," *Journal of Big Data*, vol. 11, no. 1, p. 23, Feb. 2024, doi: 10.1186/s40537-024-00882-0.

[26] L. Monje, R. A. Carrasco, and M. Sánchez-Montañés, "Machine learning XAI for early loan default prediction," *Computational Economics*, vol. 67, no. 5, pp. 4033–4062, May 2026, doi: 10.1007/s10614-025-10962-9.

BIOGRAPHIES OF AUTHORS

Ismail Idowu Akuji is an assistant lecturer in the Department of Computer Science at Abdulrasaq Abubakar Toyin University, Oke-Ogba, Kwara State. He holds a Master of Science (M.Sc.) degree in Computer Science from Kwara State University. He is a registered member of the Computer Professionals Registration Council of Nigeria (CPN). He has publications in the areas of artificial intelligence and machine learning. His research interests include data science, machine learning, and artificial intelligence. He can be contacted at email: akujiismaheel@gmail.com.

Ahmed Babajide Olanrewaju received his first degree from University of Ibadan, Nigeria, in 2006 from the Computer Science department. He also received the Master's degree from Ibadan also. He is currently the Acting Director (Head) of Information and Communication Technology (ICT) at the Federal University of Health Sciences, Ila-Orangun (FUHSI), Nigeria, and a researcher in computer science. His main research interests focus on responsible and explainable AI (XAI), Digital Transformation in Higher Education, and Geospatial analytics. He can be contacted at email: abono2000@gmail.com.



Taofik Abiodun Ahmed is currently the Director of Student Industrial Work Experience Scheme (SIWES) Kwara State College of Arabic and Islamic Legal Studies (CAILS), Ilorin. He is the current Public Relations Officer (PRO) of the Nigeria Computer Society, Kwara State Chapter. He has one of the highest and most impressive research and mentoring outputs in his field. He has been actively involved in Research, Innovation, and Development (RID) with a number of colleagues over the last 11 years, focusing on areas such as computer networks, network management, web-enabled systems, computer network performance, artificial intelligence, data analytics, education in Nigeria, and entrepreneurship. He has published numerous articles that have featured prominently in reputable national and international journals and conference proceedings. He can be contacted at email: ahmedtaophic1@gmail.com.

Ayodeji Jubril Alabi received his first degree (B.Sc.) in Computer Science from Kwara State University (KWASU), Malete, Nigeria, in 2022. He is currently pursuing his M.Sc. degree in Computer Science at Kwara State University (KWASU). His main research interests include artificial intelligence, machine learning, deep learning, computer vision and data science applications. He can be contacted at email: jubriltechguy@gmail.com.

Idris Babatunde Adeyemi holds a Bachelor of Science degree in Computer Science from Kwara State University and is currently pursuing a Master of Science in Computer Science at the University of Ilorin, where he is in the final stage of his thesis research. He is also a trained Data Scientist through AltSchool Africa. His academic and professional interests span artificial intelligence, machine learning, big data analytics, and software engineering. Idris is passionate about advancing research and developing innovative technological solutions that address contemporary challenges in computer science and related fields. He can be contacted at email: adeyemiidris818@gmail.com.