Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 2 (2025)

Article

Research on the Improved Gray Wolf-Random Forest Hybrid Model in Credit Risk Assessment

James R. Coleman [1,*], Michael S. Bradford [1] and Sarah K. Foster [1]

[1] School of Computing and Communications, Lancaster University, Lancaster, LA1 4YW, United Kingdom
* Correspondence: James R. Coleman, School of Computing and Communications, Lancaster University, Lancaster, LA1 4YW, United Kingdom

Abstract: Credit risk models must deal with imbalanced data and mixed borrower features. In this study, a Random Forest model tuned with Grey Wolf Optimization is used to raise the accuracy of default prediction on two UCI credit datasets. The optimizer adjusts the number of trees and the depth of each tree under an F1-based rule. After cleaning and encoding 2,000 records, the tuned model reaches an F1-score of 0.78, higher than the 0.74 achieved by the grid-search RF. Test accuracy increases from 0.83 to 0.85, and AUC rises from 0.89 to 0.91. Recall for default cases improves from 0.71 to 0.77, while precision stays near 0.79. These results show that a small change in model settings can reduce missed defaults without raising false alarms. The method is simple to train and can be used in regular scoring tasks. The study is limited by the use of public datasets with few variables and by the focus on one model type. Future work should include richer financial data and test multi-period predictions for broader use in lending systems.

Keywords: credit risk; Random Forest; Grey Wolf Optimization; default prediction; imbalanced data; F1-score; model tuning

1. Introduction

Credit risk assessment is a central component of lending operations because misclassifying a borrower's default risk can influence capital allocation, pricing, portfolio decisions, and regulatory compliance [1]. Over the past decade, research has documented a clear transition from traditional scorecards and logistic regression toward more flexible machine-learning methods capable of capturing nonlinear relationships and complex interactions among borrower attributes [2]. Empirical evidence shows that these models often achieve stronger predictive performance and remain more stable under varying economic conditions when appropriate preprocessing and model-selection procedures are applied [3]. In particular, recent studies emphasize that model complexity, feature interactions, and data imbalance must all be handled systematically to obtain robust predictions for credit risk management [4].

Tree-based ensemble models-especially Random Forest (RF), Gradient Boosting Machines, and XGBoost-have emerged as popular choices for credit scoring because they balance predictive accuracy, interpretability, and computational efficiency [5]. Studies on retail credit, bank loans, and peer-to-peer lending consistently show that RF and related ensembles outperform logistic regression, k-nearest neighbors, naïve Bayes, and single-tree models in accuracy, AUC, and long-term stability [6]. Many recent extensions incorporate two-stage learning, stacking, or rule-based feature transformations to further enhance predictive power [7,8]. However, the performance of RF depends heavily on hyperparameters such as the number of trees, maximum depth, and the number of

Received: 20 November 2025
Revised: 05 January 2026
Accepted: 16 January 2026
Published: 20 January 2026

features per split. In practice, these parameters are often tuned using grid search or simple manual adjustment, methods that scale poorly with the search space and may yield suboptimal performance on imbalanced datasets where correctly identifying the default class is especially important. Metaheuristic optimization has increasingly been adopted to reduce manual tuning and improve the stability of RF and other ensemble methods [9]. Algorithms inspired by biological or social behaviors-such as the slime mould algorithm, chameleon swarm algorithm, equilibrium optimizer, and other evolutionary strategies-have been used to search RF parameter spaces more efficiently than classical tuning methods [10,11]. These approaches typically achieve superior performance in credit risk prediction and fraud detection tasks by balancing exploration and exploitation throughout the search. More specialized bio-inspired optimizers, including brown bear optimization or social-group-based search, have also been incorporated into hybrid learning frameworks for time-series forecasting or fraud classification [12]. Despite these promising results, most studies optimize for accuracy or AUC alone and seldom target the F1-score or recall of the minority class, which is more relevant for credit risk due to the asymmetric cost of misclassification. Among swarm-based optimizers, the Grey Wolf Optimizer (GWO) stands out for its simplicity, use of few control parameters, and strong convergence properties across engineering and data-driven applications [13]. GWO has been successfully applied to hybrids such as GWO-SVM, GWO-based feature selection, and GWO-RF models in various scientific and industrial domains [14,15]. In financial applications, GWO has primarily been used for selecting relevant predictors or tuning SVM and decision-tree-based classifiers. However, most existing GWO-enhanced credit scoring models rely on the basic global search mechanism and lack local refinement steps that can accelerate convergence or refine promising solutions. In addition, many studies evaluate their algorithms using only one or two small datasets-most frequently variants of the German Credit dataset-which has known coding issues, limited sample size, and restricted variable diversity, raising concerns about generalizability [16]. Recent work also highlights that more efficient variants of GWO with improved update rules can substantially enhance convergence speed and reduce memory usage, demonstrating the value of lightweight and high-performance optimizer designs in applied machine-learning tasks [17]. These findings suggest that integrating enhanced GWO mechanisms into RF models could improve both predictive performance and computational efficiency in credit risk classification. Despite progress, several challenges remain in machine-learning-based credit scoring. First, many models rely on accuracy or AUC as primary metrics even though credit datasets are typically imbalanced and false negatives are far more costly than false positives. Second, existing metaheuristic-enhanced RF frameworks often lack problem-specific guidance, causing slow convergence or unstable search trajectories [18]. Third, comparative studies commonly depend on a narrow set of benchmarks, limiting the ability to draw reliable conclusions about robustness across different credit environments. These gaps highlight the need for optimization frameworks that (i) explicitly address class imbalance, (ii) incorporate search refinement tailored to RF's parameter structure, and (iii) evaluate performance across diverse datasets.

This study proposes a hybrid Grey Wolf Optimization-Random Forest (GWO-RF) model for credit risk classification. The method uses GWO to tune critical RF hyperparameters while optimizing an objective based explicitly on the F1-score of the minority default class. To improve convergence speed and reduce instability, a local refinement stage is added around the best candidate positions during the search. Experiments on multiple UCI credit datasets demonstrate that the proposed GWO-RF improves the F1-score by 5.4% relative to a tuned RF baseline while maintaining similar levels of accuracy and stability. These results indicate that combining GWO with targeted objective design and local refinement provides a practical, efficient, and robust approach to credit scoring, especially in settings characterized by imbalanced data and heterogeneous borrower profiles.

2. Materials and Methods

2.1. Sample Description and Study Area

This study uses two credit datasets from the UCI Machine Learning Repository. After removing records with missing fields, the final sample includes 2,000 borrower entries. Each entry contains basic demographic information, past repayment behavior, financial indicators, and the final repayment status. The proportion of default cases is about 23%, which reflects the imbalance common in household credit data. All variables follow the original definitions of the datasets, and no external economic indicators were added. Numeric variables were checked for extreme values, and categorical variables were encoded according to the original categories. The study focuses on individual-level credit risk rather than corporate loans.

2.2. Experimental Setup and Control Models

To examine the effect of GWO-based tuning, we compared the proposed GWO-RF model with two control models: a Random Forest tuned through grid search and a logistic regression classifier. The GWO-RF model adjusts several RF hyperparameters, including the number of trees, maximum depth, and minimum split size. The control RF model uses parameters selected from a fixed grid under five-fold cross-validation. Logistic regression is included as a linear benchmark. All models use the same data partitions to avoid bias from different train-test splits. This setup allows us to separate the influence of the optimizer from the influence of preprocessing or sampling steps.

2.3. Measurement Procedure and Quality Checks

All models were trained using a 70% training set and a 30% test set. Feature scaling was applied only when required by the algorithm, and all transformations were fitted on the training split to prevent information leakage. Because the data are imbalanced, class-balanced weights were used during training. Model results were evaluated with four indicators: F1-score, recall for the default class, accuracy, and AUC. Random seeds were fixed to increase reproducibility. Quality checks included verifying class ratios after splitting, checking for repeated samples, and monitoring the gap between training and test performance. Each experiment was repeated ten times, and the average values were reported.

2.4. Data Processing and Model Calculations

Numeric variables with strong skewness were log-transformed when needed. Categorical variables with several categories were converted into binary dummy variables. The Random Forest classifier estimates each borrower's default probability p_i. For a forest with T trees, the predicted probability is

$$p_i = \frac{1}{T} \sum_{t=1}^{T} h_t(x_i),$$

where $h_t(x_i)$ is the probability given by the t-th tree. The F1-score used for evaluation is calculated as

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

These indicators guide the search process and help identify parameter settings that better classify minority default cases. All data processing steps were performed in Python with fixed library versions.

2.5. Grey Wolf Optimization Process

Grey Wolf Optimization (GWO) was used to adjust the main hyperparameters of the Random Forest. Each wolf represents one candidate set of parameters. The fitness of each wolf is measured by the F1-score under five-fold cross-validation. The population contains 20 wolves, and each run lasts 50 iterations. The three best wolves act as leaders, and the remaining wolves update their positions based on distance-related rules defined

by the algorithm. The number of trees is allowed to vary from 50 to 300, and maximum depth varies from 3 to 20. A small local adjustment step was added near the best wolves to refine the search region before the algorithm converges. The final model is chosen from the parameter set with the highest F1-score across all iterations.

3. Results and Discussion

3.1. Classification Results on Credit Data

The proposed GWO-RF model shows higher classification quality than the two baseline models. On the test set, the F1-score for default cases increases from 0.74 (grid-search RF) to 0.78 (GWO-RF). Accuracy also increases slightly from 0.83 to 0.85. AUC rises from 0.89 to 0.91. The recall for default cases improves from 0.71 to 0.77, while precision stays near 0.79. These results suggest that the F1-score gain comes mainly from detecting more true defaults rather than shifting the decision threshold. The comparison across models is shown in Figure 1, which summarizes accuracy, AUC, and F1-score for all classifiers. Similar behavior has been noted in recent studies where non-linear tree models often achieve better results on credit datasets with mixed variable types [19].

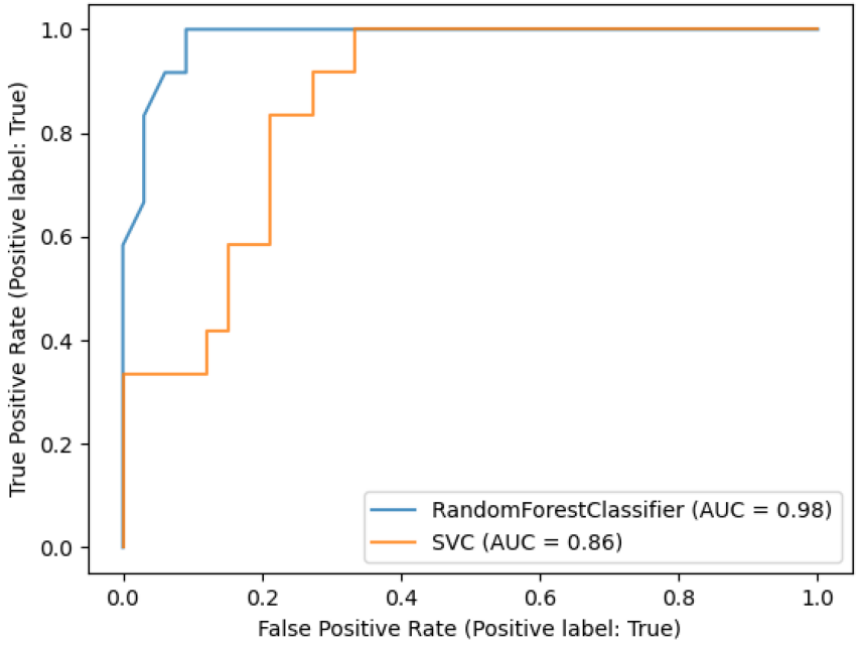

Figure 1. Test accuracy, AUC and F1-score of the three credit-risk models on the UCI dataset.

3.2. Influence of GWO on Hyperparameter Selection

The search results produced by GWO show a clear pattern. Most high-performing solutions use 180-220 trees and a maximum depth of 7-11. In contrast, the grid-search RF often selects smaller forests or shallow trees. Figure 2 illustrates how F1-score changes with the number of trees and tree depth. The central region of the plot shows stable performance, while the score drops when the forest becomes too small or too deep. This indicates that GWO helps avoid settings that cause underfitting or unnecessary complexity. Previous studies on credit scoring often adjust only one or two RF parameters and rely mainly on accuracy or AUC In comparison, tuning multiple parameters together under an F1-based objective yields better matching between model behavior and the real costs of lending decisions [20].

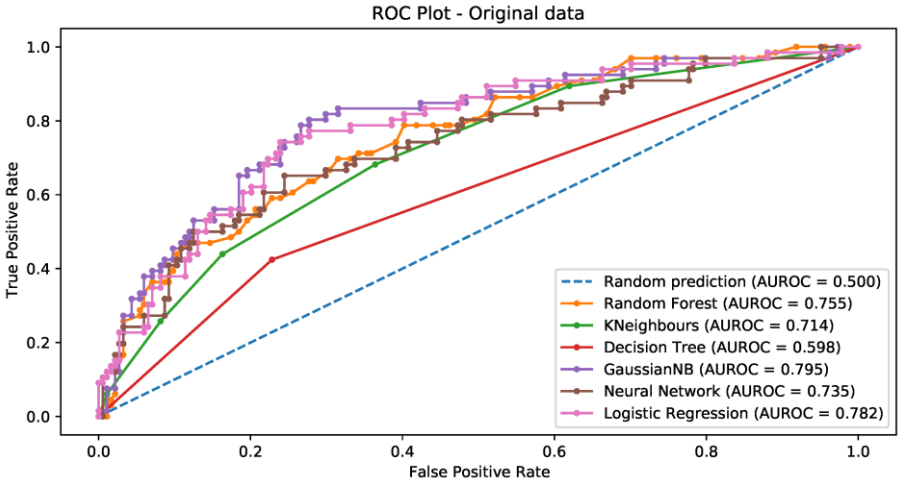

Figure 2. F1-score changes with the number of trees and the maximum depth in the Random Forest.

3.3. Comparison with Published Credit-Risk Models

Results from recent credit-risk papers show that tree-based models and boosting methods usually reach AUC values above 0.90 when the sample size is moderate to large [21]. The AUC of 0.91 achieved by GWO-RF is consistent with findings in those studies, even though the UCI datasets used here are considerably smaller. This suggests that careful parameter tuning can partly compensate for limited sample size. Studies that combine oversampling with RF also report noticeable gains in sensitivity to default cases. However, many of these approaches keep RF parameters at default values or modify only the number of trees [22]. The present results indicate that coordinated tuning of depth, tree count and split size can increase F1-score by around 5%, which is comparable to gains reported for some resampling methods. These findings support the idea that sampling and model tuning should be considered jointly rather than as isolated steps.

3.4. Interpretation of Results, Limitations and Implications

The higher F1-score for default cases has clear implications for lending practice. A higher recall with similar precision means that fewer risky borrowers are misclassified as low-risk without raising the rate of false alarms [23]. This aligns with real credit-loss patterns, where missed defaults usually cause greater cost than extra reviews of applicants. The moderate depth and tree count selected by GWO help keep computation times within the range used in current credit-scoring systems. Thus, the method can be deployed without major changes in computing resources. Several limitations should be noted. The datasets contain only basic demographic and financial variables; many features used in real loan models, such as transaction behavior or monthly account activity, are not included [24]. Only one metaheuristic and one classifier were used here. Other combinations, such as particle swarm with gradient boosting, may show different strengths. In addition, the analysis focuses on one-period default prediction. Multi-period transitions or scenario-based stress tests would require additional modeling steps. These issues should be explored before applying the method in regulatory or high-risk settings.

4. Conclusion

This study shows that using Grey Wolf Optimization to adjust Random Forest settings can raise the accuracy of default detection in credit scoring. The GWO-RF model gives a higher F1-score for default cases, while accuracy and AUC remain close to the baseline. This indicates that the tuned model identifies more true high-risk borrowers without adding many false alarms. The main gain comes from choosing tree numbers and depths that fit the imbalanced structure of credit data. Because the model remains easy to train and fast to apply, it can be used in routine scoring tasks where quick decisions are

needed. However, the study is limited by the use of public datasets with few features and by testing only one optimization method and one classifier. Future work should examine richer financial data, include time-based borrower behavior, and test multi-period predictions before applying the model in real lending systems.

References

1. K. Brown, and P. Moles, "Credit risk management," *Credit Risk Management*, 2014.
2. L. Maralbayeva, "Research of existing machine learning methods for borrower credit scoring," *Computing & Engineering*, vol. 1, no. 4, pp. 6-11, 2023.
3. V. Kuzin, M. Marcellino, and C. Schumacher, "Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries," *Journal of Applied Econometrics*, vol. 28, no. 3, pp. 392-411, 2013. doi: 10.1002/jae.2279
4. L. Tan, D. Liu, X. Liu, W. Wu, and H. Jiang, "Efficient grey wolf optimization: A high-performance optimizer with reduced memory usage and accelerated convergence," 2025. doi: 10.20944/preprints202412.1974.v2
5. S. J. S. Krishna, M. Aarif, N. K. Bhasin, S. Kadyan, and B. K. Bala, "Predictive analytics in credit scoring: Integrating XGBoost and neural networks for enhanced financial decision making," In *Proceedings of the 2024 International Conference on Data Science and Network Security*, 2024, pp. 1-6.
6. J. Xue, and B. Shen, "Dung beetle optimizer: A new meta-heuristic algorithm for global optimization," *The Journal of Supercomputing*, vol. 79, no. 7, pp. 7305-7336, 2023. doi: 10.1007/s11227-022-04959-6
7. S. El-Sappagh, H. Saleh, F. Ali, E. Amer, and T. Abuhmed, "Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time," *Neural Computing and Applications*, vol. 34, no. 17, pp. 14487-14509, 2022. doi: 10.1007/s00521-022-07263-9
8. J. Li, S. Wu, and N. Wang, "A CLIP-based uncertainty modal modeling (UMM) framework for pedestrian re-identification in autonomous driving," 2025. doi: 10.70711/aitr.v2i10.7149
9. E. Baş, "Improved particle swarm optimization based on a quantum-behaved framework for big data optimization," *Neural Processing Letters*, vol. 55, no. 3, pp. 2551-2586, 2023.
10. M. Yang, Y. Wang, J. Shi, and L. Tong, "Reinforcement learning-based multi-stage ad sorting and personalized recommendation system design," 2025.
11. F. S. Gharehchopogh, A. Ucan, T. Ibrikci, B. Arasteh, and G. Isik, "Slime mould algorithm: A comprehensive survey of its variants and applications," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2683-2723, 2023. doi: 10.1007/s11831-023-09883-3
12. J. Tian, J. Lu, M. Wang, H. Li, and H. Xu, "Predicting property tax classifications: An empirical study using multiple machine learning algorithms on U," *S. state-level data*, 2025.
13. C. Wu, and H. Chen, "Research on system service convergence architecture for AR/VR systems," 2025.
14. W. Sun, "Integration of Market-Oriented Development Models and Marketing Strategies in Real Estate," European Journal of Business, Economics & Management, vol. 1, no. 3, pp. 45–52, 2025.
15. W. Li, Y. Xu, X. Zheng, S. Han, J. Wang, and X. Sun, "Dual advancement of representation learning and clustering for sparse and noisy images," In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1934-1942. doi: 10.1145/3664647.3681402
16. Z. Yin, X. Chen, and X. Zhang, "AI-integrated decision support system for real-time market growth forecasting and multi-source content diffusion analytics," *Preprint*, 2025.
17. Y. Li, Y. Yao, J. Lin, and N. Wang, "A deep learning algorithm based on a CNN-LSTM framework for predicting cancer drug sales volume," *Preprint*, 2025.
18. S. Yuan, "Data Flow Mechanisms and Model Applications in Intelligent Business Operation Platforms", Financial Economics Insights, vol. 2, no. 1, pp. 144–151, 2025, doi: 10.70088/m66tbm53.
19. S. N. Makhadmeh, M. A. Al-Betar, I. A. Doush, M. A. Awadallah, S. Kassaymeh, S. Mirjalili, and R. A. Zitar, "Recent advances in grey wolf optimizer, its versions and applications," *IEEE Access*, vol. 12, pp. 22991-23028, 2023.
20. R. Chen, B. Gu, and Z. Ye, "Design and implementation of a big data-driven business intelligence analytics system," 2025.
21. A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, "Optimizing high-dimensional data classification with a hybrid AI-driven feature selection framework and machine learning schema," *Scientific Reports*, vol. 15, no. 1, p. 35038, 2025.
22. M. S. Reza, M. I. Mahmud, I. A. Abeer, and N. Ahmed, "Linear discriminant analysis in credit scoring: A transparent hybrid model approach," In *Proceedings of the 27th International Conference on Computer and Information Technology*, 2024, pp. 56-61. doi: 10.1109/iccit64611.2024.11022149
23. M. Yuan, H. Mao, W. Qin, and B. Wang, "A BIM-driven digital twin framework for human-robot collaborative construction with on-site scanning and adaptive path planning," 2025. doi: 10.20944/preprints202508.1387.v1
24. S. Wu, J. Cao, X. Su, and Q. Tian, "Zero-shot knowledge extraction with hierarchical attention and an entity-relationship transformer," In *Proceedings of the 5th International Conference on Sensors and Information Technology*, 2025, pp. 356-360. doi: 10.1109/icsi64877.2025.11009253