Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 8 No. 3, November 2023, pp. 363 – 375
DOI:

Enhancing Default Prediction in P2P Lending using Random Forest and Grey Wolf Optimization-based Feature Selection

Bagus Winarko Nugroho[1], **Purwanto**[2]
[1,2]Universitas Dian Nuswantoro, Universitas Dian Nuswantoro
E-mail : P31202002371@mhs.dinus.ac.id[1], purwanto@dsn.dinus.ac.id

Heribertus Himawan[3]
[3]Universitas Dian Nuswantoro
E-mail : himawan26@dsn.dinus.ac.id

Abstract - Online lending services such as Peer to Peer (P2P) loans provide convenience for lenders to transact directly without involving banks as intermediaries. Identifying potential loan recipients who are at risk of default is a crucial step in preventing financial losses, as lenders are responsible for default risk. However, predicting default risk becomes a challenge when P2P lending datasets have various complex features. Some features in P2P lending are redundant, while others do not significantly contribute to an effective solution. Therefore, feature selection is an important process to choose a relevant subset of features from input or target data. Traditional feature selection methods often fail to provide optimal results. A better approach is to use heuristic search algorithms capable of finding suboptimal feature subsets. We employ the Grey Wolf Optimization (GWO) technique, inspired by the hierarchy of leadership and grey wolf hunting mechanisms. Combined with Random Forest (RF), which has limitations in classifying data with very high dimensions, our GWO+RF combination has proven to enhance classification performance better than previous research. It achieves an accuracy score of 97.31%, compared to previous research with scores of only 67.72% for RBM+RF, 64% for Binary PSO+ERT, and 92% for GA+RF.

Keywords – P2P lending; Feature selection; Grey Wolf Optimization; Random Forest; Accuracy

1. INTRODUCTION

P2P lending has the potential to bring about significant changes to the trajectory of traditional banks in the future. Being the world's largest digital credit marketplace, P2P lending offers various types of loans, including personal, business, and medical loans. The primary objective of P2P lending when it emerged in 2005 was to democratize access to more efficient consumer financial services. This approach involves individuals presenting loan proposals, which are then approved by investors or lenders, bypassing the need for a formal financial institution's involvement. The inception of P2P lending occurred in the UK with Zopa in 2005. Subsequently, in 2006, the United States introduced LendingClub and Prosper, while China established its own credit platforms.

Numerous researchers have examined P2P lending through the introduction of a diverse array of models. This includes the utilization of a stacking ensemble model for the assessment of personal credit risk[1]. Logistic regression, hinged on credit scores, and linear regression, based on profit scores, have been employed to forecast the likelihood of default and potential profits within a novel loan recommendation framework[2]. In terms of default prediction, three statistical models (Logistic Regression (LR), Bayesian Classifier, and Linear Discriminant Analysis (LDA)) along with five AI models (Decision Tree, Random Forest, Light-GBM, Artificial Neural Network (ANN), and Convolutional Neural Network (CNN)) have been utilized[3]. A multi-view deep neural network has been crafted to address default prediction challenges inherent in imbalanced and intricate datasets[4]. Within the realm of credit risk scoring, there exists a benchmarking model grounded in machine learning techniques[5]. The forecasting of default risk on unbalanced datasets has been approached via three models: Random Forest, Neural Network, and Logistic Regression[6]. Additionally, an innovative resampling ensemble model, rooted in data distribution, has been proposed for the assessment of credit risk within imbalanced datasets[7]. The exploration of multi-view ensemble learning, incorporating the distance-to-model concept and adaptive clustering, has been carried out for credit risk assessment within imbalanced datasets[8]. Predictive models, encompassing Naive Bayes, Decision Tree, and Boosted Decision Tree methodologies, have been employed for default prediction[9][10]. Credit risk assessment has been facilitated through Regression models[11]. A multi-round ensemble learning model, leveraging heterogeneous ensemble frameworks, has been developed for the prediction of defaults[12]. ANN-based models have been harnessed to categorize credit scores for both default and non-default classifications in the context of P2P lending[13]. Malekipirbazari and Aksakalli have implemented Random Forest (RF) for the identification of top-tier borrowers, contrasting them with FICO credit scores derived from LC scores[14]. Abnormal investor identification and the prediction of potential investors have been rooted in outlier detection utilizing poor credit scores[15]. Lastly, the utilization of the internal rate of return has been employed to anticipate the anticipated profits for investors[16].

P2P lending datasets, extracted from P2P lending platforms, frequently contain irrelevant or redundant features. Consequently, the predictive performance of models often falls short of optimal levels, yielding inaccurate outcomes[17]. The extensive size and complexity of P2P lending datasets lead to suboptimal and inefficient model performance[18]. For instance, the processing duration tends to elongate due to the extensive feature processing required[19]. To mitigate these issues, feature selection emerges as a solution. This method discerns pertinent features, those exerting significant influence on the prediction process. Moreover, it serves to curtail data dimensions and eliminate irrelevant features, thereby enhancing classification model accuracy[20].

Feature selection methods tailored for P2P loan default prediction have been extensively proposed. The Max-Relevance and Min-Redundancy (MRMR) technique is employed for feature selection, and k-means clustering aids in discarding irrelevant attributes[1]. Various models, including LightGBM (Light Gradient Boosting Machine)[21], Random Forest[22][23], Logistic Regression[24], Random Forest and XGBoost[24], Binary Particle Swarm Optimization (PSO) with Support Vector Machines (SVM)[25], Adaptive Feature Selection based on Most Informative Graph and Most Relative Graph[26], and Grey Relational Clustering, have demonstrated the superior accuracy of results achieved through feature selection compared to unselected features[27].

However, traditional feature selection methods face limitations in addressing the challenge. The transition from a set of different N-sized features to a 2^N-sized vector exacerbates the already vast feature space[28]. The evaluation of 2^N subsets falls under the category of np-hard problems[20][22]. As an alternative solution for such complexities, Evolutionary Algorithms (EA) step in. The Gray Wolf Optimization (GWO), introduced by Mirjalili, Mirjalili, and Lewis, is a type of EA well-suited for optimizing feature selection within extensive feature spaces[29]. GWO incorporates a hierarchical leadership concept and the hunting mechanism of gray wolves, making it effective for addressing optimization problems. Therefore, this paper introduces a novel feature selection approach founded on the GWO technique and Machine Learning models. The objective is to achieve highly accurate loan default prediction for P2P lending, utilizing data collected from P2P lending platforms. What sets it apart from previous research is that prior studies employed less effective feature selection algorithms when dealing with high-dimensional feature datasets. Therefore, the GWO algorithm combined with Random Forest classification serves as the optimal solution for enhancing the evaluation performance, as we will discuss in the results section.

2. RESEARCH METHOD

Figure 1 illustrates the process flow of the research methodology, specifically the employment of Gray Wolf Optimization for Feature Selection in the prediction of defaults within the context of P2P lending.

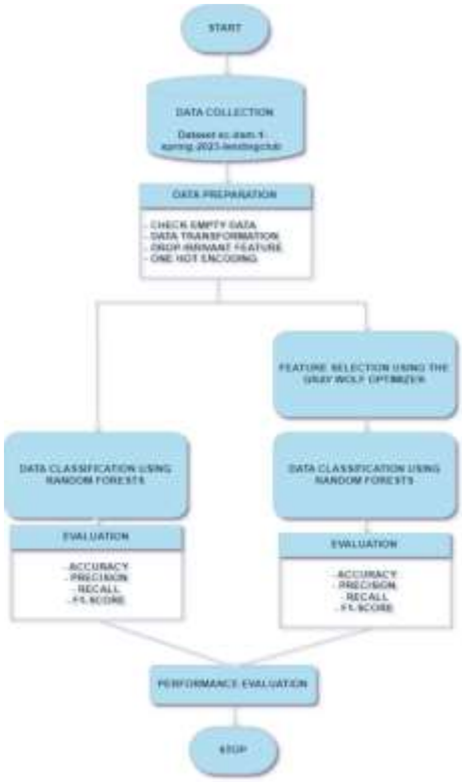

Figure 1. The Workflow of research method

2.1. P2P lending dataset

The research dataset is sourced from the Kaggle repository, a dataset of online P2P loans available at the following link: https://www.kaggle.com/competitions/ec-dsm-1-spring-2023-lendingclub/data in the year 2023. The P2P loan dataset comprises 16,000 records with 29 features, as depicted in Figure 2.

2.2. Data Preparation

2.2.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is employed to detect absent data, trends, and anomalies within P2P lending datasets. The objective of EDA is to grasp the pertinent attributes influencing the dependent variable. EDA also contributes to enhancing the model's predictive capability. This procedural approach aligns with data preprocessing to detect and rectify data inconsistencies effectively. Figure 3 illustrates the findings from the exploration and analysis of data, unveiling the correlations among numerical attributes within the P2P Lending dataset.

2.2.2. Pre-Processing Data

The objective of preprocessing the data involves amalgamating, refining, and minimizing the dataset. The Sklearn library is utilized to detect any missing values, and all features within the P2P lending dataset exhibit no missing values. Features indicating outliers, specifically 'grade' and 'sub grade,' are excluded. Subsequently, One-Hot Encoding is applied to object data types such as term, home ownership, purpose, verification status, and payment plan. The conclusive form of the P2P lending dataset, prepared for analysis, is illustrated in Figure 4. It comprises 41 independent features, with 'default' serving as the target variable.

```
RangeIndex: 16000 entries, 0 to 15999
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   loan_amnt                16000 non-null  int64
 1   term                     16000 non-null  object
 2   grade                    16000 non-null  int64
 3   sub_grade                16000 non-null  int64
 4   emp_length               16000 non-null  int64
 5   home_ownership           16000 non-null  object
 6   purpose                  16000 non-null  object
 7   annual_inc               16000 non-null  float64
 8   dti                      16000 non-null  float64
 9   verification_status      16000 non-null  object
 10  delinq_2yrs              16000 non-null  int64
 11  fico_range_low           16000 non-null  int64
 12  inq_last_6mths           16000 non-null  int64
 13  open_acc                 16000 non-null  int64
 14  pub_rec                  16000 non-null  int64
 15  revol_bal                16000 non-null  int64
 16  revol_util               16000 non-null  float64
 17  total_acc                16000 non-null  int64
 18  installment              16000 non-null  float64
 19  int_rate                 16000 non-null  float64
 20  funded_amnt              16000 non-null  int64
 21  out_prncp                16000 non-null  float64
 22  last_pymnt_amnt          16000 non-null  float64
 23  total_rec_int            16000 non-null  float64
 24  total_rec_late_fee       16000 non-null  float64
 25  pymnt_plan               16000 non-null  object
 26  recoveries               16000 non-null  float64
 27  collection_recovery_fee  16000 non-null  float64
 28  default                  16000 non-null  int64
```

Figure 2. Feature of P2P Lending dataset original

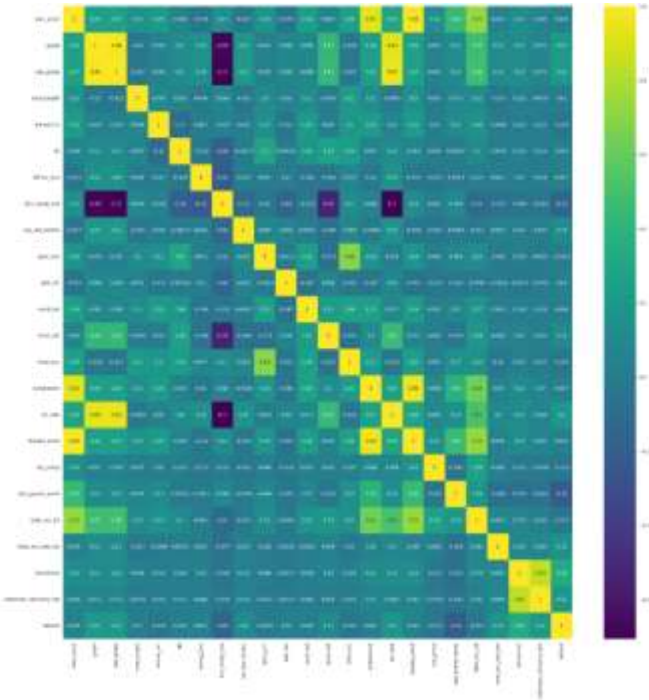

Figure 3. Lending dataset corelation matrix

```
RangeIndex: 16000 entries, 0 to 15999
Data columns (total 42 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   loan_amnt                          16000 non-null  int64
 1   term                               16000 non-null  int64
 2   emp_length                         16000 non-null  int64
 3   annual_inc                         16000 non-null  float64
 4   dti                                16000 non-null  float64
 5   delinq_2yrs                        16000 non-null  int64
 6   fico_range_low                     16000 non-null  int64
 7   inq_last_6mths                     16000 non-null  int64
 8   open_acc                           16000 non-null  int64
 9   pub_rec                            16000 non-null  int64
 10  revol_bal                          16000 non-null  int64
 11  revol_util                         16000 non-null  float64
 12  total_acc                          16000 non-null  int64
 13  installment                        16000 non-null  float64
 14  int_rate                           16000 non-null  float64
 15  funded_amnt                        16000 non-null  int64
 16  out_prncp                          16000 non-null  float64
 17  last_pymnt_amnt                    16000 non-null  float64
 18  total_rec_int                      16000 non-null  float64
 19  total_rec_late_fee                 16000 non-null  float64
 20  recoveries                         16000 non-null  float64
 21  collection_recovery_fee            16000 non-null  float64
 22  default                            16000 non-null  int64
 23  verification_status_Source Verified  16000 non-null  int64
 24  verification_status_Verified       16000 non-null  int64
 25  purpose_credit_card                16000 non-null  int64
 26  purpose_debt_consolidation         16000 non-null  int64
 27  purpose_educational                16000 non-null  int64
 28  purpose_home_improvement           16000 non-null  int64
 29  purpose_house                      16000 non-null  int64
 30  purpose_major_purchase             16000 non-null  int64
 31  purpose_medical                    16000 non-null  int64
 32  purpose_moving                     16000 non-null  int64
 33  purpose_other                      16000 non-null  int64
 34  purpose_renewable_energy           16000 non-null  int64
 35  purpose_small_business             16000 non-null  int64
 36  purpose_vacation                   16000 non-null  int64
 37  purpose_wedding                    16000 non-null  int64
 38  pymnt_plan_y                       16000 non-null  int64
 39  OTHER                              16000 non-null  int64
 40  OWN                                16000 non-null  int64
 41  RENT                               16000 non-null  int64
```

Figure 4. Dataset P2P Lending after pre-processing

2.3. *Proposed Model*

Mirjalili, Mirjalili, and Lewis introduced an approach to Evolutionary Algorithms (EA) that draws inspiration from social hierarchies. This strategy, known as Gray Wolf Optimization (GWO), is influenced by the hunting behavior of gray wolves. Gray wolves are recognized as apex predators within the food chain, typically residing in packs with an average size of 5-12 individuals.

• Alpha (α): The leader of the wolf pack, either male or female, responsible for crucial decisions like hunting, sleeping, and setting the pack's schedule.

• Beta (β): Subordinates who assist the α in making decisions and other tasks. β could be male or female, and often emerges as a potential successor to α.

• Delta (δ): These individuals submit to α and β but have authority over ϖ. Roles like scouts, sentinels, elders, hunters, and caretakers fall within this category.

• Omega (ϖ): Acts as a scapegoat and must obey the commands of other pack members. ϖ represents the lowest-ranking and weakest wolf.

Beyond the hierarchical structure exhibited by the Gray Wolf, the phenomenon of group hunting constitutes a captivating social behavior within this species. The primary stages of the Gray Wolf's hunting process encompass tracking, pursuit, and assault of prey. Both the hunting techniques of Gray Wolves and their social hierarchy are subjected to mathematical modeling to construct the GWO and carry out optimization procedures. Within the mathematical model of GWO, α is denoted as the most potent solution, β as the second-best solution, and ϖ as the third-best solution. The remaining solution candidates are treated as ϖ. During the hunt, guidance is provided by α, β, and δ, while ϖ follows these three contenders. Consequently, the herd surrounds the prey before initiating the hunt. To express this circular behavior mathematically, the following equations (1)-(2) are employed.

$$\vec{X}(t+1) = \vec{X}_p(t) + \vec{A} \cdot \vec{D} \tag{1}$$

$$\vec{D} = \left|\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)\right| \tag{2}$$

Within this context, D ⃗ is established according to equation (2), where 't' signifies the total iterations. Coefficient vectors A ⃗ and C ⃗ come into play, while X ⃗p denotes the prey's location and X ⃗ represents the Grey wolf's position. The specific values of A ⃗ and C ⃗ are ascertained through the application of equations; (3) and (4).

$$;\vec{A} = 2a \cdot \vec{r}_1 - a \tag{3}$$

$$\vec{C} = 2\vec{r}_2 \tag{4}$$

Here, the value of 'a' gradually diminishes in a linear fashion from 2 to 0 throughout the iteration process. Vectors ⃗r1 and ⃗r2 are randomly generated with values within the range of [0, 1], thereby facilitating the update of the Grey wolf's position as outlined in Equation (5).

ith X⃗α, X⃗β, and X⃗δ representing the initial three most optimal solutions within the group during a specific iteration. A⃗1, A⃗2, and A⃗3 are defined following the equation in (3). D⃗ α, D⃗ β, and D⃗ δ are established based on the equation in (7).

$$\vec{X}(t+1) = \frac{(\vec{X}_1 + \vec{X}_2 + \vec{X}_3)}{3} \tag{5}$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot \vec{D}_\alpha$$
$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta \tag{6}$$
$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot \vec{D}_\delta$$

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}|$$
$$\vec{D}_\beta = |\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}| \tag{7}$$
$$\vec{D}_\delta = |\vec{C}_3 \cdot \vec{X}_\delta - \vec{X}|$$

where C⃗1, C⃗2, C⃗3 are defined in Eq. (4)

Hence, the Gray Wolf Optimization (GWO) involves a revision of the 'a' parameter, which governs the balance between exploration and exploitation. This parameter is subject to linear updating in each iteration, ranging from 2 to 0, as outlined in Equation (8).

$$a = 2 - t\frac{2}{\max iter} \tag{8}$$

max iter denotes the maximum permissible number of iterations for the optimization process.

Algorithm 1: Pseudo code of the classical grey wolf optimization algorithm

Initialize the prey wolf population X_i $(i = 1,2, \cdots, n)$
Initialize a, $\vec{A}$ and $\vec{C}$
Calculate the fitness of each search agent
$\vec{X}_\alpha$ is the best search agent
$\vec{X}_\beta$ is the second best search agent
$\vec{X}_\delta$ is third best agent
while $(t < \max iter)$
    for each search agent
        update the position of the current search agent by Eq. (15)
    end for
    Update $a, \vec{A}, \vec{C}$
    a, $\vec{A}$ and $\vec{C}$
    Update $\vec{X}_\alpha, \vec{X}_\beta, \vec{X}_\delta$
    t=t+1
end while
return $\vec{X}_\alpha$

This paper proposes an effective feature selection method, namely GWO-RF (Grey Wolf - Random Forest), to predict default in P2P lending. There are two main phases in using GWO-RF. First, redundant and irrelevant features based on Figure 2 are eliminated by seeking the best features using GWO. GWO generates population initialization, and then the population's positions are updated in the discrete search space. Second, the GWO-RF models are executed for the classification process based on the optimal feature set obtained in the first phase. Figure 5 illustrates the workflow of the GWO-RF model.

GWO efficiently exploits the feature space to identify the optimal features within the P2P lending dataset. The optimal feature is the solution that yields the highest classification accuracy with the chosen attributes. Typically, the number of features is reduced compared to the original dataset. The fitness function defined in Equation 9 is employed to maximize the accuracy performance of the ML model and leverages GWO for assessing the selected features.

$$Fitness\ (t) = \alpha\rho + \beta\ \frac{N - L}{L} \qquad (9)$$

Where ρ represents the classification accuracy of the Random Forest Model. L denotes the length of the chosen feature set, N stands for the total count of features within the P2P lending dataset. Meanwhile, α and β are parameters associated with the weight of classification accuracy and the feature selection quality. Here, α ranges within [0, 1] and β = 1 − α. 't' signifies the iteration.

Figure 6 visually demonstrates a subset of features that correspond to potential solutions. In this context, we adopt the binary chromosome model for configuring feature subsets. 'd' denotes the total count of features, equivalent to the chromosome's length. The values '1' or '0' are assigned based on the chromosome's position. If the value of the i^{th} bit is 1, the corresponding feature is included; otherwise, if the i^{th} bit is 0, the feature is excluded.

2.3.1. Performance Evaluation

The assessment of the proposed approach's effectiveness involves employing a confusion matrix for a binary classification scenario. This matrix encompasses metrics like true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP). Utilizing the confusion matrix facilitates the computation of accuracy, recall, precision, and F1-score, which are defined as:

$$Accuracy = \frac{TP+TN}{TN+FP+TP+FN} \qquad (10)$$

$$Recall = \frac{TP}{TP+FN} \qquad (11)$$

$$Precison = \frac{TP}{TP+FP} \qquad (12)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recal} \qquad (13)$$

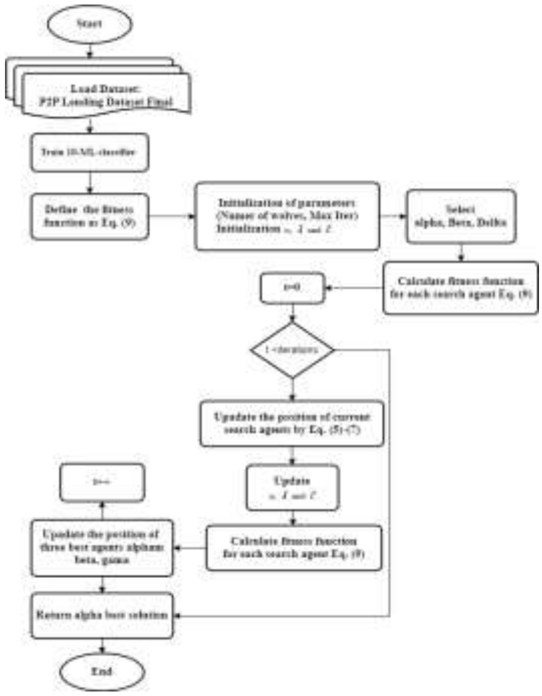

Figure 5. The workflow of GWO-RF

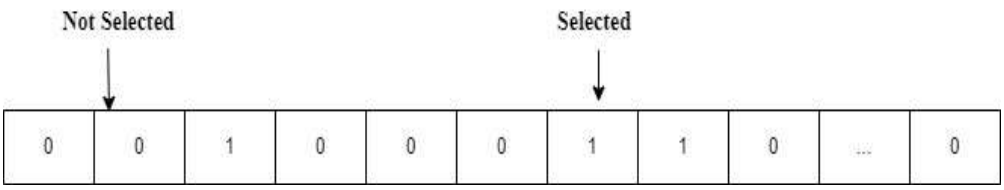

Figure 6. Solution representation of feature selection

3. RESULTS AND DISCUSSION

The outcome of the feature selection process employing GWO along with 3-ML models using the P2P lending dataset as input yields a compilation of the most influential attributes. The configuration of parameter values encompasses elements such as the iteration count, the quantity of wolves, the number of features or dimensions, the scope of the search domain, as well as alpha and beta parameters for the fitness function. These specifics are elaborated in Table 1. Executing the proposed GWO-RF model follows the outlined procedure in Figure 5, producing the fitness function output computed according to Equation (9).

As depicted in Figure 7, the fitness value consistently rises with each iteration. GWO exhibits the capability to yield the optimal solution or the most fitting attributes for the GWO-RF model classification procedure. The count of selected features (L), representing the optimal features based on the GWO-RF model, is visualized in table. This particular feature wields significant influence over default prediction within P2P lending.

Table 1. The setting of parameter for the proposed method.

Parameter	Numbers
Number of iteration	10
Number of wolves	95
Search Domain	[0 1]
Number of Dimensions	41

Alpha in fitness function	0.4999
Beta in fitness function	0.5

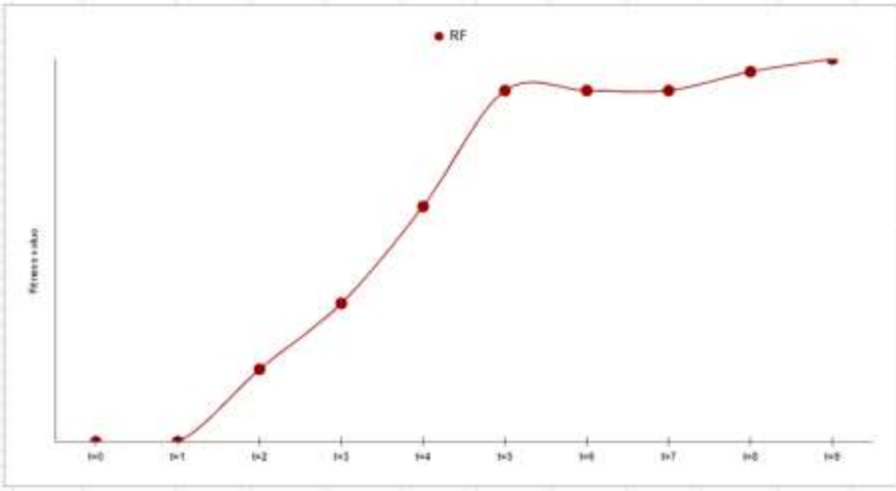

Figure 7. The fitness value of the proposed model.

Table 2. The selected features in P2P Lending.

Models	Order of features																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RF	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	
RF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

From Table 2, it is known that the features selected by the Grey Wolf Optimizer algorithm are Loan amount, term, last_pymnt_amnt, total_rec_int, and recoveries. Out of a total of 41 existing features, the most influential ones, after going through the GWO algorithm processing, are narrowed down to just 5 features. These 5 features are then processed using the Random Forest algorithm for classification.

Table 3. Performance evaluation comparison between GWO-RF and original RF.

Model	Accuracy (%)		Recall (%)		Precision (%)		F1-score (%)	
	Original	GWO	Original	GWO	Original	GWO	Original	GWO
RF	96.53	97.31	76.83	85.39	100	96.24	86.89	90.49

Based on the results from Table 3, a performance comparison between GWO-RF and the original RF is conducted. GWO-RF outperforms the Original RF based on performance evaluations for Accuracy, recall, and F1-score. Feature selection using GWO effectively improves predictive performance against the Random Forest model. For Accuracy, it increases by 0.78%, Recall by 8.56%, and F1-score by 3.6%, while Precision experiences a decrease of 3.76%.

Furthermore, Table 4 presents a comparison of the proposed model's accuracy performance with previous related studies. This table confirms that the proposed model outperforms the research conducted by Nguyen et al .[30], which used the Restricted Boltzmann Machine (RBM) feature selection algorithm in conjunction with six Machine Learning models: LDA, LR, ANN, KNN, SVM, RF, with the highest score being 81.20% for the RBM+LDA mode. Subsequently, Setiawan, Suharjito, and Diana[25] employed the Hybrid Binary PSO+ERT in predicting P2P Lending defaults, resulting in a score of 64%. Finally, Victor and Raheem [31] used GA as a feature selection method along with three Machine Learning models: LR, RF, and SVM, with the highest score achieved by GA+RF at 92% accuracy. In contrast, the model proposed in this study, GWO+RF, obtained a score of 97.31%, which represents the highest accuracy score in predicting P2P Lending defaults. Therefore, GWO has proven to be a suitable feature selection optimization algorithm. However, there is a need to expand it through an extended search space to accommodate high-dimensional datasets.

Table 4. Comparison between the model proposed in this study and previous related research.

Study	Feature selection	Models	Accuracy (%)
Nguyen et al. [30]	Restricted Boltzmann Machine	LDA	81.20
		LR	81.05
		ANN	66.05
		KNN	72.55
		SVM	76.56
		RF	67.72
Setiawan, Suharjito and Diana [25]	Binary Particle Swarm Optimisation	ERT	64
Victor and Raheem [31]	Genetic Algorithm	LR	86
		RF	92
		SVM	85
Proposed model	**GWO**	**RF**	**97.31**

4. CONCLUSION

The conclusion that can be drawn from the evaluation of the Grey Wolf Optimizer-Random Forest method is the feature selection approach proposed in this research. The GWO-RF model can select relevant features and disregard irrelevant ones within the P2P Lending dataset. Comparative studies of three performance evaluations (accuracy, recall, and F1-score) indicate that the GWO-RF model outperforms the original Random Forest method in predicting defaults in P2P Lending. The proposed method is also superior to three previous related studies based on accuracy. Furthermore, there is a need to enhance GWO by expanding the search space to handle high-dimensional datasets.

REFERENCES

[1] W. Yin, B. Kirkulak-Uludag, D. Zhu, and Z. Zhou, "Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending," *Appl. Soft Comput.*, vol. 142, 2023, doi: 10.1016/j.asoc.2023.110302.

[2] Y. Rong, S. Liu, S. Yan, W. W. Huang, and Y. Chen, "Proposing a new loan recommendation framework for loan allocation strategies in online P2P lending," *Ind. Manag. Data Syst.*, vol. 123, no. 3, pp. 910–930, 2023, doi: 10.1108/IMDS-07-2022-0399.

[3] P. C. Ko, P. C. Lin, H. T. Do, and Y. F. Huang, "P2P Lending Default Prediction Based on AI and Statistical Models," *Entropy*, vol. 24, no. 6, 2022, doi: 10.3390/e24060801.

[4] Y. Tan and G. Zhao, "Multi-view representation learning with Kolmogorov-Smirnov to predict default based on imbalanced and complex dataset," *Inf. Sci. (Ny).*, vol. 596, pp. 380–394, 2022, doi: 10.1016/j.ins.2022.03.022.

[5] V. Moscato, A. Picariello, and G. Sperlí, "A benchmark of machine learning approaches

for credit score prediction," *Expert Syst. Appl.*, vol. 165, 2021, doi: 10.1016/j.eswa.2020.113986.
[6] Y. R. Chen, J. S. Leu, S. A. Huang, J. T. Wang, and J. I. Takada, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," *IEEE Access*, vol. 9, pp. 73103–73109, 2021, doi: 10.1109/ACCESS.2021.3079701.
[7] K. Niu, Z. Zhang, Y. Liu, and R. Li, "Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending," *Inf. Sci. (Ny).*, vol. 536, pp. 120–134, 2020, doi: 10.1016/j.ins.2020.05.040.
[8] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending," *Inf. Sci. (Ny).*, vol. 525, pp. 182–204, 2020, doi: 10.1016/j.ins.2020.03.027.
[9] M. J. Christ, R. N. P. Tri, W. Chandra, and T. Mauritsius, "Lending club default prediction using Naïve Bayes and decision tree," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 2528–2534, 2019, doi: 10.30534/ijatcse/2019/99852019.
[10] A. Semiu and A. A. R. Gilal, "A boosted decision tree model for predicting loan default in P2P lending communities," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 1257–1261, 2019, doi: 10.35940/ijeat.A9626.109119.
[11] S. F. Chen, G. Charkaborty, L. H. Li, and C. T. Lin, "Credit risk assessment using regression model on P2P lending," *Int. J. Appl. Sci. Eng.*, vol. 16, no. 2, pp. 149–157, 2019.
[12] W. Li, S. Ding, H. Wang, Y. Chen, and S. Yang, "Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China," *World Wide Web*, vol. 23, no. 1, pp. 23–45, 2020, doi: 10.1007/s11280-019-00676-y.
[13] A. Byanjankar, M. Heikkila, and J. Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," *Proc. - 2015 IEEE Symp. Ser. Comput. Intell. SSCI 2015*, pp. 719–725, 2015, doi: 10.1109/SSCI.2015.109.
[14] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, 2015, doi: 10.1016/j.eswa.2015.02.001.
[15] H. Li, Y. Zhang, N. Zhang, and H. Jia, "Detecting the Abnormal Lenders from P2P Lending Data," *Procedia Comput. Sci.*, vol. 91, pp. 357–361, 2016, doi: 10.1016/j.procs.2016.07.095.
[16] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," *Decis. Support Syst.*, vol. 89, pp. 113–122, 2016, doi: 10.1016/j.dss.2016.06.014.
[17] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.
[18] H. D. Wang, "Research on the features of car insurance data based on machine learning," *Procedia Comput. Sci.*, vol. 166, pp. 582–587, 2020, doi: 10.1016/j.procs.2020.02.016.
[19] M. Papoušková and P. Hájek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decis. Support Syst.*, vol. 118, no. October 2018, pp. 33–45, 2019.
[20] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Comput.*, vol. 22, no. 3, pp. 811–822, 2018, doi: 10.1007/s00500-016-2385-6.
[21] M. A. Muslim *et al.*, "New model combination meta-learner to improve accuracy

prediction P2P lending with stacking ensemble learning," *Intell. Syst. with Appl.*, vol. 18, 2023, doi: 10.1016/j.iswa.2023.200204.

[22] A. K. Sharma, L. H. Li, and R. Ahmad, "Default Risk Prediction Using Random Forest and XGBoosting Classifier," *Smart Innov. Syst. Technol.*, vol. 314, pp. 91–101, 2023, doi: 10.1007/978-3-031-05491-4_10.

[23] J. Xu, D. Chen, and M. Chau, "Identifying features for detecting fraudulent loan requests on P2P platforms," *IEEE Int. Conf. Intell. Secur. Informatics Cybersecurity Big Data, ISI 2016*, pp. 79–84, 2016, doi: 10.1109/ISI.2016.7745447.

[24] S. F. Chen, G. Chakraborty, and L. H. Li, "Feature Selection on Credit Risk Prediction for Peer-to-Peer Lending," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11717 LNAI, pp. 5–18, 2019, doi: 10.1007/978-3-030-31605-1_1.

[25] N. Setiawan, Suharjito, and Diana, "A comparison of prediction methods for credit default on peer to peer lending using machine learning," *Procedia Comput. Sci.*, vol. 157, pp. 38–45, 2019, doi: 10.1016/j.procs.2019.08.139.

[26] L. Cui, Y. Jiao, L. Bai, L. Rossi, and E. R. Hancock, "Adaptive feature selection based on the most informative graph-based features," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10310 LNCS, pp. 276–287, 2017, doi: 10.1007/978-3-319-58961-9_25.

[27] L. He, H. Xu, and G. Y. Ke, "A hybrid predictive framework for evaluating P2P credit risks," *Grey Syst.*, vol. 12, no. 3, pp. 551–573, 2022, doi: 10.1108/GS-03-2021-0041.

[28] C. Shen and K. Zhang, "Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional classification," *Complex Intell. Syst.*, vol. 8, no. 4, pp. 2769–2789, 2022, doi: 10.1007/s40747-021-00452-4.

[29] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[30] T. Nguyen Truong, S. Khuat Thanh, T. Ngo Thi Thu, N. Nguyen Ha, and D. Tran Manh, "Improve Risk Prediction in Online Lending (P2P) Using Feature Selection and Deep Learning," *Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 11, pp. 216–222, 2019.

[31] L. Victor and M. Raheem, "Loan Default Prediction Using Genetic Algorithm: A Study within Peer-To-Peer Lending Communities," *Int. J. Innov. Sci. Res. Technol.*, vol. 6, no. 3, pp. 1195–1205, 2021.