Metaheuristic-Driven Feature Selection for Machine Learning-Based Credit Scoring

Chinh Xuan Pham
Department of Information Technology, Banking Academy of Viet Nam, Hanoi, Vietnam
chinhpx@hvnh.edu.vn (corresponding author)

Huynh Ngoc Trinh
Institute for Artificial Intelligence, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
huynhtn@vnu.edu.vn

Long Quoc Tran
Institute for Artificial Intelligence, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
tqlong@vnu.edu.vn

Received: 18 October 2025 | Revised: 16 November 2025 | Accepted: 3 December 2025

 DOI: https://doi.org/10.48084/etasr.15590

ABSTRACT

The recurrence of financial and debt crises in recent years has underscored the critical importance of effective credit risk management in financial research. Within this context, credit scoring is a critical tool for financial institutions to evaluate loan applications and has received extensive attention in both statistical and machine learning research. This study proposes a novel credit scoring framework that integrates metaheuristic-driven feature selection with machine learning classifiers. Three metaheuristic algorithms, namely Particle Swarm Optimization (PSO), Simulated Annealing (SA), and Tabu Search (TS), are employed to identify the most relevant subset of features, whereas a wide range of machine learning models is trained to determine the most effective combination for credit scoring. The framework is evaluated on three benchmark datasets, namely Australian, German, and Japanese datasets, from the UCI Machine Learning Repository. Experimental results show that metaheuristic-based feature selection consistently improves model performance compared to the baseline without feature selection and conventional methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Stepwise Selection, demonstrating its effectiveness and robustness in credit scoring tasks.

Keywords-credit scoring; feature selection; metaheuristic search; machine learning

I. INTRODUCTION

Credit borrowing refers to the practice whereby individuals or organizations obtain funds or financial resources in advance from banks or financial institutions, with the obligation to repay later, usually with interest [1]. Before approving credit, lenders commonly apply credit scoring to evaluate a customer's reliability and support lending decisions, making credit-related activities essential to nearly all forms and scales of financial transactions [2, 3]. Typically, credit scoring is conducted based on a customer's historical transactions and financial information [4]. However, traditional manual statistical analysis has currently become inadequate in today's financial systems, where big data with their vast, complex, and unstructured characteristics pose major challenges to deriving valuable insights [5].

Recent advancements in Artificial Intelligence (AI) have enabled models to capture complex relationships in data and demonstrate robustness when dealing with large-scale information. Therefore, AI is increasingly leveraged to enhance credit scoring, providing more accurate, efficient, and dynamic evaluations of a customer's creditworthiness [6, 7]. Many studies have investigated approaches to improve credit scoring, particularly through machine learning [8-10] and deep learning [11-13]. Machine learning is valued for its efficiency, transparency, and strong ability to handle structured financial data, making it highly suitable for real-world lending decisions [14-16]. Although deep learning offers powerful predictive capabilities, it generally requires much larger datasets and computational resources and is often less interpretable [17, 18]. Moreover, the rise of Explainable AI (XAI) provides interpretable and transparent outputs, which is increasingly recognized as both essential and transformative for credit

scoring models [19, 20]. These benefits are particularly prominent in machine learning, which can further strengthen trust, fairness, and adoption in lending practices.

Additionally, a significant issue in credit scoring is that databases often contain numerous borrower-related variables, where excessive or irrelevant information may introduce noise into the dataset [21, 22]. Such redundancy not only complicates the learning process and degrades model performance but also increases computational costs during training and testing. To address this, feature selection serves as an effective strategy to streamline the dataset by identifying the most informative variables, thereby improving efficiency and predictive accuracy [23, 24]. Traditional feature selection methods are commonly classified into filter, wrapper, and embedded approaches [25]. Filter methods are computationally efficient but tend to overlook feature interdependencies, whereas wrapper and embedded techniques generally achieve higher accuracy at the cost of increased computational complexity, especially with high-dimensional data [26, 27]. Among recent advances, a prominent and promising group of methods for feature selection is metaheuristic-based optimization, which has attracted considerable attention due to its flexibility and global search capability [28-33]. Inspired by natural or social behaviors, these algorithms efficiently explore the search space to identify near-optimal feature subsets, offering a better balance between global exploration and local exploitation in complex and nonlinear problems. More recently, some studies have explored combining metaheuristic feature selection with quantum machine learning models, such as Variational Quantum Classifiers, to enhance predictive performance in financial applications like credit card fraud detection [34]. Despite these advances, most existing studies remain largely isolated, focusing on individual algorithms or specific datasets, and only a few systematically compare these novel combinations across multiple datasets, highlighting a clear research gap that the present study aims to address.

In this paper, we develop a comprehensive machine learning pipeline for the credit scoring problem. We evaluate multiple machine learning models to assess their performance on the task. Furthermore, we address the feature selection problem using metaheuristic search methods to identify the most informative input data. Three metaheuristic algorithms, including Particle Swarm Optimization (PSO), Simulated Annealing (SA), and Tabu Search (TS), are integrated into the pipeline to compare their effectiveness and determine the best configuration for credit scoring. Three benchmark datasets are employed in the experiments, evaluated using five standard performance metrics: Accuracy (Acc), F1-score (F1), Area Under the Curve (AUC), Brier Score (BS), and Kolmogorov–Smirnov (KS) statistic. The experimental results demonstrate that the proposed approaches, which incorporate metaheuristic-driven feature selection, consistently outperform their counterparts without feature selection as well as traditional dimensionality reduction methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Stepwise Selection. These findings highlight the effectiveness and suitability of integrating metaheuristic optimization for enhanced credit scoring performance.

II. RESEARCH METHODOLOGY

In this section, we develop a comprehensive pipeline for the credit scoring problem. The proposed pipeline integrates machine learning models to predict credit scores and assist financial organizations in making lending decisions. Furthermore, we incorporate a metaheuristic-driven feature selection approach to identify the most important variables in large and redundant datasets, thereby enhancing predictive accuracy and reducing computational cost. The pipeline is designed to flexibly integrate various machine learning and metaheuristic methods, allowing for extensive experimentation and identification of the most effective model for the task.

A. Pipeline

Our pipeline provides a comprehensive framework for addressing the credit scoring problem, as illustrated in Figure 1. The initial steps follow conventional procedures, including data splitting and preprocessing to prepare the dataset for modeling. The core component of our pipeline is the metaheuristic-driven feature selection stage, which plays a crucial role in eliminating irrelevant features, reducing data dimensionality, and enhancing both model performance and interpretability. Additionally, the baseline scenario (None, as shown in Figure 1) represents the case without feature selection and is included for ablation study comparisons. Subsequently, multiple machine learning models are evaluated across five performance metrics to identify the most effective approach for this problem. Overall, the entire pipeline is designed around integrating learning models with metaheuristic-driven feature selection methods to improve the overall performance of credit scoring.

B. Metaheuristic-Driven Feature Selection

Our approach focuses on feature selection driven by metaheuristic search methods for the credit scoring problem, aiming to identify the most informative subset of features while reducing redundancy and noise. The process begins with GridSearchCV, a well-established technique widely adopted in previous studies [35, 36], to fine-tune the estimator's hyperparameters and ensure optimal baseline performance. Once the estimator is optimized, it serves as the evaluation function to compute performance metrics (e.g., accuracy, F1-score, or error rate) for each candidate feature subset generated during the search process.

The metaheuristic algorithms then iteratively explore the feature space by carefully balancing exploration, which broadly searches across possible subsets, and exploitation, which refines promising solutions. During each iteration, candidate solutions are updated based on their performance, and the process continues until convergence or a stopping criterion is met. Finally, the optimal feature subset is selected and used as input features for the downstream machine learning models in the credit scoring task. Figure 2 illustrates the overall workflow of the metaheuristic-driven feature selection process, highlighting the integration of estimator tuning, evaluation, and iterative optimization. In the following sections, we provide a detailed discussion of each specific metaheuristic method employed for feature selection, including PSO [37], SA [38], and TS [39].

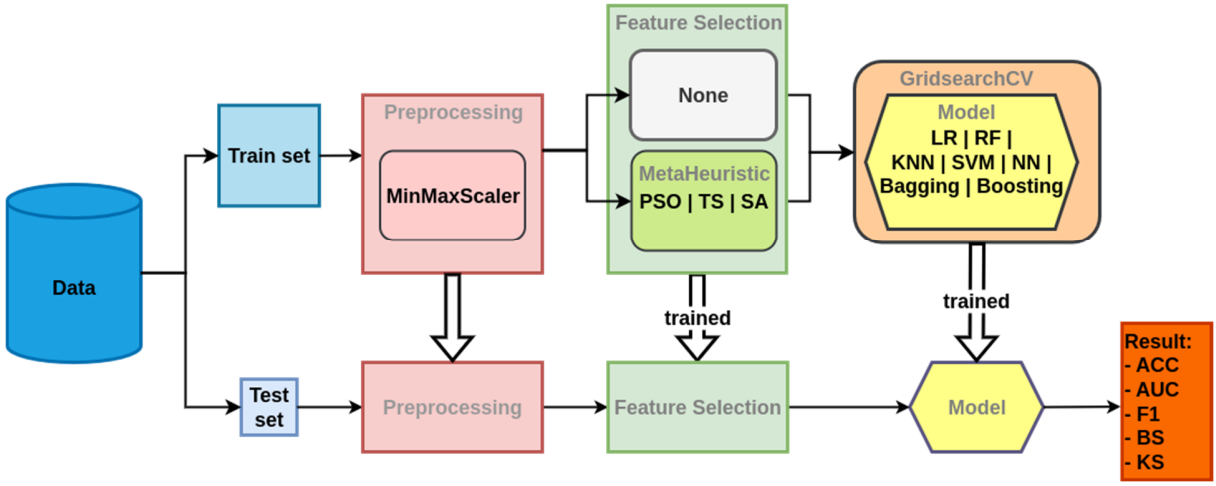

Fig. 1. Proposed credit scoring pipeline: data preparation, preprocessing, feature selection, and model training and evaluation.

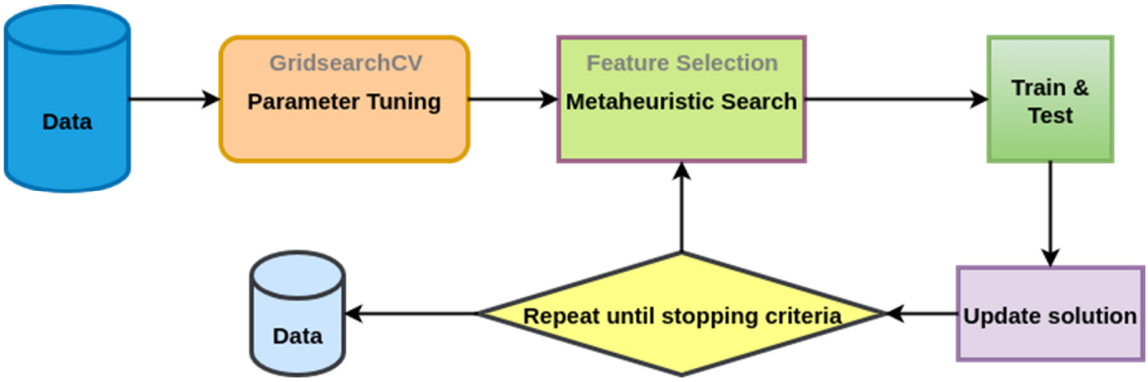

Fig. 2. Proposed metaheuristic-based feature selection pipeline.

1) Particle Swarm Optimization

Algorithm 1: PSO for Feature Selection
Input: Dataset D, n, T, v_{max}
Output: Optimal feature subset
Initialize population $x[i] \in \{0,1\}^n$
Initialize velocity $v[i] \in [-v_{max}, v_{max}]^n$
For each particle:
    Evaluate $score(x[i], D)$
    $p\_best[i] \leftarrow x[i]$
$g\_best \leftarrow best(p\_best[\,])$
For $t = 1$ to T do:
    For each particle:
        $update\_v(v[i], p\_best[i], g\_best)$
        $update\_x(x[i], v[i])$
        Evaluate $score(x[i], D)$
        If $score(x[i], D) > score(p\_best[i], D)$:
            $p\_best[i] \leftarrow x[i]$
    $g\_best \leftarrow best(p\_best[\,])$
Return g_best

The PSO [37] algorithm is inspired by the collective behavior of organisms in nature (called particles), such as bird flocking or fish schooling. Initially, it was proposed to simulate social behavior and later applied to solve global optimization problems. Importantly, PSO can also be effectively applied to feature selection, as shown in Algorithm 1.

This method is fitted on a dataset D, where n denotes the number of input features, and iterates for T iterations to find the optimal feature subset through the global best solution g_best. The algorithm begins by creating a fixed number of particles $x[\,]$, each representing a candidate feature subset encoded as a binary vector, where each bit takes the value 0 (feature not selected) or 1 (feature selected) along with a corresponding personal best position $p\_best[]$. Each particle $x[i]$ is associated with a velocity $v[i]$ that determines how its position evolves in subsequent iterations, initialized within the range $[-v_{max}, v_{max}]$.

In each iteration, the velocity $v[i]$ and position $x[i]$ are updated toward its own best-known position $p\_best[i]$ and the global best position g_best, according to the standard PSO equations [37]. Each updated particle is then re-evaluated using a classifier on dataset D to measure its performance based on the currently selected feature subset. The personal best position $p\_best[i]$ of a particle is also updated if its new solution performs better than its previous best. Finally, the global best position g_best is updated by comparing all $p\_best[\,]$ values across the swarm.

2) Simulated Annealing

Algorithm 2: SA for Feature Selection
Input: Dataset D, n, T_{max}, T_{min}, α
Output: Optimal feature subset
Initialize solution $x \in \{0,1\}^n$
Evaluate $score(x, D)$
$g\_best \leftarrow x$; $T \leftarrow T_{max}$
While $T > T_{min}$:
    $y \leftarrow x$
    $Random\_flip(y)$
    Evaluate $score(y, D)$
    $\Delta = score(y, D) - score(x, D)$
    If $(\Delta > 0)$:
        $x \leftarrow y$
    Else If $(exp(\Delta / T) > rand(0,1))$:
        $x \leftarrow y$
    If $\big(score(x, D) > score(g\_best, D)\big)$:
        $g\_best \leftarrow x$
    $T \leftarrow T \times \alpha$
Return g_best

The SA [38] algorithm is a probabilistic metaheuristic inspired by the physical process of annealing in metallurgy, where a material is gradually cooled to reach a stable, low-energy state. Its application to feature selection is shown in Algorithm 2. In this context, a binary vector $x \in \{0,1\}^n$ encodes which features are selected. The algorithm starts by evaluating an initial subset on dataset D and storing it as the global best solution g_best.

At each iteration, a neighbor y is generated by flipping one or more bits of x. If the new subset improves the score ($\Delta > 0$), it replaces the current one. Otherwise, it may still be accepted with probability $exp(\Delta / T)$ where T is the temperature that controls the chance of accepting worse solutions. The temperature decreases gradually following a cooling schedule $T \leftarrow T \times \alpha$, with $0 < \alpha < 1$ as the cooling rate that determines how quickly the search shifts from exploration to exploitation.

By allowing occasional acceptance of inferior moves at high temperatures and becoming more selective as T cools, SA avoids premature convergence and steadily refines toward an optimal and compact feature subset after repeated iterations between T_{max} and T_{min}.

3) Tabu Search

Algorithm 3: Tabu Search for Feature Selection
Input: Dataset D, n, T, K, tenure
Output: Optimal feature subset
Initialize solution $x \in \{0,1\}^n$
Evaluate $score(x, D)$
$g\_best \leftarrow x$
Initialize $tabu\_last[\,] \leftarrow (-1)$
For $t = 1$ to T do:
    $C \leftarrow \emptyset$
    $J \leftarrow Sample(\{1, \ldots, n\}, K)$
    For each $j \in J$:
        $y \leftarrow x$; $y[j] \leftarrow (1 - y[j])$
        Evaluate $score(y, D)$
        If $(t - tabu\_last[j] > tensure)$ or $(score(y, D) > score(g\_best, D))$:
            $C \leftarrow C \cup \{y, j\}$
    If $C = \emptyset$:
        $j^* \leftarrow argmin\_\{j\}(tabu\_last)$
        $y^* \leftarrow x$; $y^*[j^*] \leftarrow (1 - y^*[j^*])$
        Evaluate $score(y^*, D)$
    Else:
        Evaluate each solution in C
        $(y^*, j^*) \leftarrow argmax\_\{y, j\}(C)$
    $x \leftarrow y^*$; $tabu\_last[j^*] \leftarrow t$
    If $(score(x, D) > score(g\_best, D))$:
        $g\_best \leftarrow x$
Return g_best

The TS [39] algorithm is another metaheuristic optimization technique designed to overcome the limitations of local search methods by preventing cycling and enabling exploration beyond local optima. Its use for feature selection is described in Algorithm 3. In this context, TS iteratively modifies a binary vector $x \in \{0,1\}^n$, where each bit indicates whether a feature is selected. The algorithm begins by evaluating the initial subset on dataset D and storing it as the global best solution g_best.

In each iteration, TS generates a set of neighboring solutions by flipping K randomly chosen feature bits. To prevent reversing recent moves, it maintains a $tabu\_last[\,]$ array, which records the last iteration when each feature was modified. A move is forbidden (tabu) for a fixed number of iterations called the tenure. However, if the new candidate y achieves a better score than the current best g_best, it is accepted according to the aspiration criterion, even if tabu.

After evaluating all admissible neighbors, the best candidate y^* replaces the current solution, and its flipped feature index j^* is updated in the array $tabu\_last$. The global best g_best is also updated if y^* outperforms previous bests. By controlling the tenure value, TS balances diversification (exploring new regions) and intensification (refining good areas), leading to an optimal feature subset after T iterations.

III. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed approach on financial datasets related to the credit scoring problem. First, we provide a summary of each dataset to highlight their key characteristics and differences. Next, we report the baseline performance of the machine learning models in addressing this task. Finally, we investigate the impact of metaheuristic-based feature selection on improving overall predictive performance and learning efficiency, and compare it with other feature selection methods to demonstrate its effectiveness and practicality.

A. Datasets

All experiments in this study were performed on three benchmark credit scoring datasets sourced from the UCI Machine Learning Repository, which offers a diverse and comprehensive foundation to ensure the robustness and reliability of our analysis. The Australian Credit Approval (Australian) dataset [40] is one of the most commonly utilized benchmarks in credit scoring studies. It contains 690 samples characterized by 14 input variables representing credit card applications. Another widely used dataset is the German Credit Data (German) dataset [41], which categorizes applicants as good or bad credit risks based on multiple attributes. In this work, we adopted the numerical-only version of this dataset, comprising 1,000 instances with 24 variables. The third dataset, the Japanese Credit Screening (Japanese) dataset [42], includes 689 records described by 15 variables. The key characteristics of all three datasets are summarized in Table I.

TABLE I. SUMMARIZATION OF THE THREE DATASETS

Dataset	No. of instances	Default / not default	No. of features
Australian	690	383 / 307	14
German	1,000	300 / 700	24
Japanese	689	383 / 306	15

B. Base Results (without Feature Selection)

Initially, we provide a general evaluation of the performance of various machine learning models across the credit scoring datasets. Ten models, ranging from simple classifiers to complex ensemble methods, were employed to identify the best-performing approach using five common classification metrics, as presented in Table II. Overall, the metrics did not exhibit significant differences across models on the three datasets. For the Australian dataset, the Logistic Regression (LR) model achieved the highest performance in terms of AUC (0.8715), F1-score (0.8575), and KS (0.7430), whereas the Bagging model yielded the best Accuracy (0.8699) and Brier Score (0.1301), though with only marginal improvement. On the German dataset, the XGBoost model consistently achieved the best results across all five metrics, with an Accuracy of 0.7674 and an AUC of 0.6859, indicating that this dataset is more challenging than the Australian one. For the Japanese dataset, the Gradient Boosting Decision Trees (GBDT) model attained the best Accuracy (0.8716), AUC (0.8725), Brier Score (0.1284), and KS (0.7450), whereas the Multilayer Perceptron (MLP) model achieved the highest F1-score (0.8619). In general, the best and most stable performance was achieved by ensemble models such as Bagging, XGBoost, and GBDT, which combine multiple learners to produce more robust and reliable predictions for credit scoring tasks.

TABLE II. BASELINE RESULTS ON THREE DATASETS

Models	Metrics														
	Australian					German					Japanese				
	Acc	AUC	F1	BS	KS	Acc	AUC	F1	BS	KS	Acc	AUC	F1	BS	KS
AdaBoost	0.8633	0.8631	0.8455	0.1367	0.7261	0.7526	0.6791	0.5398	0.2474	0.3581	0.8485	0.8502	0.8376	0.1515	0.7003
Bagging	**0.8699**	0.8692	0.8506	**0.1301**	0.7384	0.7610	0.6812	0.5456	0.2390	0.3624	0.8689	0.8697	0.8576	0.1311	0.7394
ExtraTree	0.8595	0.8592	0.8380	0.1405	0.7183	0.7534	0.6553	0.4952	0.2466	0.3106	0.8579	0.8577	0.8432	0.1421	0.7153
GBDT	0.8601	0.8605	0.8432	0.1399	0.7210	0.7674	0.6740	0.5259	**0.2326**	0.3480	**0.8716**	**0.8725**	0.8596	**0.1284**	**0.7450**
KNN	0.8460	0.8469	0.8327	0.1540	0.6938	0.7314	0.6185	0.4262	0.2686	0.2370	0.8561	0.8602	0.8463	0.1439	0.7204
LR	0.8691	**0.8715**	**0.8575**	0.1309	**0.7430**	0.7666	0.6832	0.5469	0.2334	0.3663	0.8558	0.8614	0.8543	0.1442	0.7228
MLP	0.8543	0.8544	0.8379	0.1457	0.7089	0.7574	0.6743	0.5333	0.2426	0.3486	0.8668	0.8708	**0.8619**	0.1332	0.7417
RF	0.8642	0.8616	0.8456	0.1358	0.7232	0.7552	0.6497	0.4843	0.2448	0.2995	0.8695	0.8690	0.8584	0.1305	0.7379
SVM	0.8574	0.8574	0.8419	0.1457	0.7148	0.7572	0.6740	0.5324	0.2428	0.3479	0.8646	0.8717	0.8603	0.1354	0.7435
XGBoost	0.8624	0.8608	0.8417	0.1376	0.7216	**0.7674**	**0.6859**	**0.5542**	**0.2326**	**0.3718**	0.8643	0.8642	0.8535	0.1357	0.7284

C. Results with Metaheuristic-Driven Feature Selection

To demonstrate the effectiveness of metaheuristic-based feature selection, we conducted additional experiments using three methods, PSO, SA, and TS, on the three credit scoring datasets. We then compared their performance with the baseline models without feature selection. In addition, our approach was evaluated through mutual comparison and against traditional feature selection techniques, including PCA retaining 90% of the variance, LDA, and Stepwise Selection. These experiments provide a comprehensive analysis that underscores the effectiveness and advantages of the proposed method.

Most evaluation metrics showed notable improvements when integrating metaheuristic-driven feature selection, outperforming the baseline results presented in Table II across all three datasets. Firstly, the combination of machine learning models with PSO achieved the most consistent performance gains, as shown in Table III. The improvement was most significant on the Australian dataset, where the best model (LR) surpassed the baseline across all metrics, achieving an Accuracy of 0.8832 (an increase of 1.33%), a KS value of 0.7694 (an increase of 0.0264), and a Brier Score of 0.1168 (a decrease of 0.0133). In contrast, the Japanese dataset exhibited moderate improvement, with the best model (RF) showing a 1.01% increase in Accuracy, whereas the German dataset achieved a smaller gain of approximately 0.6% in Accuracy compared with the best corresponding baseline results. Secondly, consistent improvements were also observed when integrating SA-driven feature selection with each machine learning model across the three datasets, as presented in Table IV. On the Australian dataset, the combination with RF achieved the best overall performance, outperforming the baseline across all five metrics, specifically, the F1-score reached 0.8721, representing an improvement of 0.0146 compared with the best baseline result. Similarly, on the

German dataset, the integration with GBDT yielded performance gains, although the improvements were less pronounced due to the challenging nature of this dataset. For the Japanese dataset, the Bagging model delivered the best performance, with a KS value of 0.7686, marking an improvement of 0.0236 over the best baseline result. Lastly, the incorporation of TS-based feature selection further enhanced the performance of the models, as shown in Table V. The XGBoost model achieved the best performance across four metrics, except for F1-score, where LR performed best. All top results under the TS-based approach showed improvements over their corresponding baselines. Similarly, the German and Japanese datasets also exhibited performance gains across most machine learning models, with the best results obtained by Support Vector Machine (SVM) and GBDT, respectively.

TABLE III. PROPOSED METHOD: PSO RESULTS ON THREE DATASETS

Models	**Metrics**														
	Australian					**German**					**Japanese**				
	Acc	**AUC**	**F1**	**BS**	**KS**	**Acc**	**AUC**	**F1**	**BS**	**KS**	**Acc**	**AUC**	**F1**	**BS**	**KS**
AdaBoost	0.8405	0.8396	0.8283	0.1595	0.6792	0.7592	0.6729	0.5249	0.2408	0.3458	0.8610	0.8650	0.8510	0.1390	0.7300
Bagging	0.8393	0.8389	0.8229	0.1607	0.6778	0.7520	0.6753	0.5404	0.2480	0.3506	0.8695	0.8674	0.8546	0.1305	0.7349
ExtraTree	0.8497	0.8495	0.8360	0.1503	0.6989	0.7472	0.6542	0.5013	0.2528	0.3085	0.8707	0.8690	0.8548	0.1293	0.7380
GBDT	0.8601	0.8598	0.8459	0.1399	0.7195	0.7536	0.6779	0.5430	0.2464	0.3558	0.8561	0.8537	0.8370	0.1439	0.7074
KNN	0.8520	0.8486	0.8202	0.1480	0.6972	0.7224	0.6301	0.4639	0.2776	0.2601	0.8622	0.8621	0.8509	0.1378	0.7243
LR	**0.8832**	**0.8847**	**0.8693**	**0.1168**	**0.7694**	0.7464	0.6586	0.5092	0.2536	0.3173	0.8524	0.8587	0.8492	0.1476	0.7174
MLP	0.8543	0.8582	0.8434	0.1457	0.7164	0.6776	0.5077	0.0938	0.3224	0.0349	0.8695	0.8733	0.8651	0.1305	0.7466
RF	0.8740	0.8706	0.8570	0.1260	0.7411	0.7544	0.6563	0.4983	0.2456	0.3125	**0.8817**	**0.8823**	**0.8769**	**0.1183**	**0.7645**
SVM	0.8555	0.8573	0.8408	0.1445	0.7147	0.7568	0.6734	0.5235	0.2432	0.3468	0.8573	0.8591	0.8530	0.1427	0.7183
XGBoost	0.8509	0.8489	0.8272	0.1491	0.6977	**0.7680**	**0.6828**	**0.5445**	**0.2320**	**0.3657**	0.8780	0.8771	0.8701	0.1220	0.7542

TABLE IV. PROPOSED METHOD: SA RESULTS ON THREE DATASETS

Models	**Metrics**														
	Australian					**German**					**Japanese**				
	Acc	**AUC**	**F1**	**BS**	**KS**	**Acc**	**AUC**	**F1**	**BS**	**KS**	**Acc**	**AUC**	**F1**	**BS**	**KS**
AdaBoost	0.8474	0.8470	0.8359	0.1526	0.6940	0.7360	0.6396	0.4750	0.2640	0.2793	0.8707	0.8785	0.8616	0.1293	0.7569
Bagging	0.8462	0.8476	0.8333	0.1538	0.6951	0.7360	0.6587	0.5150	0.2640	0.3174	**0.8829**	**0.8843**	0.8686	**0.1171**	**0.7686**
ExtraTree	0.8520	0.8529	0.8278	0.1480	0.7057	0.7288	0.6395	0.4786	0.2712	0.2789	0.8537	0.8553	0.8404	0.1463	0.7105
GBDT	0.8335	0.8342	0.8070	0.1665	0.6685	**0.7680**	**0.6882**	**0.5489**	**0.2320**	**0.3764**	0.8561	0.8585	0.8544	0.1439	0.7170
KNN	0.8486	0.8499	0.8406	0.1514	0.6998	0.7368	0.6296	0.4509	0.2632	0.2592	0.8561	0.8592	0.8417	0.1439	0.7184
LR	0.8740	0.8763	0.8658	0.1260	0.7525	0.7240	0.6229	0.4450	0.2760	0.2458	0.8671	0.8739	0.8606	0.1329	0.7479
MLP	0.8694	0.8730	0.8633	0.1306	0.7459	0.7296	0.5417	0.1639	0.2704	0.0834	0.8183	0.8153	0.8058	0.1817	0.6307
RF	**0.8775**	**0.8775**	**0.8721**	**0.1225**	**0.7549**	0.7624	0.6584	0.4994	0.2376	0.3167	0.8610	0.8629	0.8440	0.1390	0.7258
SVM	0.8613	0.8649	0.8492	0.1387	0.7298	0.7456	0.6414	0.4634	0.2544	0.2829	0.8720	0.8771	**0.8699**	0.1280	0.7543
XGBoost	0.8451	0.8435	0.8310	0.1549	0.6871	0.7536	0.6584	0.5002	0.2464	0.3168	0.8537	0.8557	0.8407	0.1463	0.7115

TABLE V. PROPOSED METHOD: TS RESULTS ON THREE DATASETS

Models	**Metrics**														
	Australian					**German**					**Japanese**				
	Acc	**AUC**	**F1**	**BS**	**KS**	**Acc**	**AUC**	**F1**	**BS**	**KS**	**Acc**	**AUC**	**F1**	**BS**	**KS**
AdaBoost	0.8243	0.8229	0.8066	**0.1757**	0.6458	0.7424	0.6519	0.4951	0.2576	0.3037	0.8598	0.8668	0.8541	0.1402	0.7335
Bagging	0.8566	0.8558	0.8437	0.1434	0.7116	0.7336	0.6452	0.4836	0.2664	0.2905	0.8451	0.8451	0.8287	0.1549	0.6901
ExtraTree	0.8601	0.8576	0.8303	0.1399	0.7153	0.7384	0.6535	0.5047	0.2616	0.3070	0.8817	0.8816	0.8747	0.1183	0.7631
GBDT	0.8624	0.8622	0.8512	0.1376	0.7244	0.7488	**0.6715**	**0.5358**	0.2512	**0.3430**	**0.8854**	0.8848	0.8737	**0.1146**	0.7695
KNN	0.8601	0.8588	0.8411	0.1399	0.7177	0.7120	0.6044	0.4059	0.2880	0.2088	0.8488	0.8543	0.8410	0.1512	0.7085
LR	0.8671	0.8669	**0.8578**	0.1329	0.7338	0.7496	0.6614	0.5102	0.2504	0.3227	0.8671	0.8767	0.8517	0.1329	0.7533
MLP	0.8705	0.8697	0.8504	0.1295	0.7395	0.6128	0.4999	0.1294	0.3872	0.0205	0.7829	0.7705	0.6810	0.2171	0.5410
RF	0.8717	0.8714	0.8552	0.1283	0.7428	0.7528	0.6522	0.4915	0.2472	0.3044	0.8500	0.8524	0.8426	0.1500	0.7047
SVM	0.8428	0.8513	0.8358	0.1572	0.7025	**0.7544**	0.6497	0.4845	**0.2456**	0.2994	0.8817	**0.8892**	**0.8788**	0.1183	**0.7783**
XGBoost	**0.8740**	**0.8750**	0.8537	**0.1260**	**0.7499**	0.7496	0.6592	0.5052	0.2504	0.3183	0.8585	0.8592	0.8467	0.1415	0.7185

Furthermore, we provide a comprehensive evaluation comparing three metaheuristic methods both mutually and against three traditional methods, including PCA (0.9), LDA, and Stepwise Selection. The best results across ten machine learning models are selected and presented in Figures 3, 4, and 5 for full comparison. Overall, the metaheuristic approaches consistently outperform the traditional methods, alternately achieving the best performance across all three datasets, with the most stable results from the SA method. For the Australian dataset, the Accuracy comparison is shown in Figure 3. The PSO combined with LR achieves the highest Accuracy of 0.8832, outperforming all three traditional feature selection methods, the best among which (PCA) is still 1.33% lower. Meanwhile, the other two metaheuristic methods, SA and TS,

also yield better results than the traditional ones. For the German dataset (Figure 4), the Brier Score metric is reported, where smaller values indicate better performance. Both SA and PSO achieve the best results, surpassing all three traditional methods. However, TS performs relatively worse, remaining higher than PCA and LDA, though still better than Stepwise Selection. Finally, for the Japanese dataset (Figure 5), the KS index obtained by TS shows a remarkable improvement, significantly outperforming traditional methods such as PCA and LDA (0.0249 and 0.0232, respectively). SA also achieves better results than the traditional approaches, whereas PSO performs slightly below Stepwise Selection.

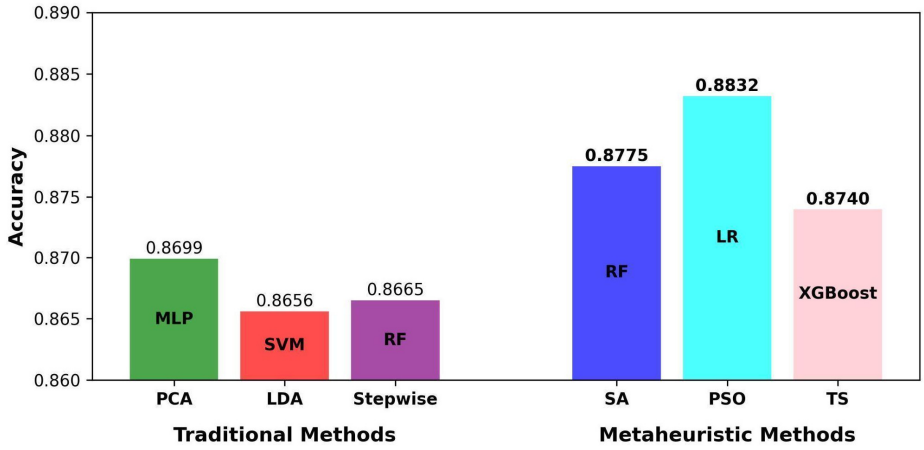

Fig. 3. Accuracy comparison of metaheuristic and traditional feature selection methods on the Australian dataset.

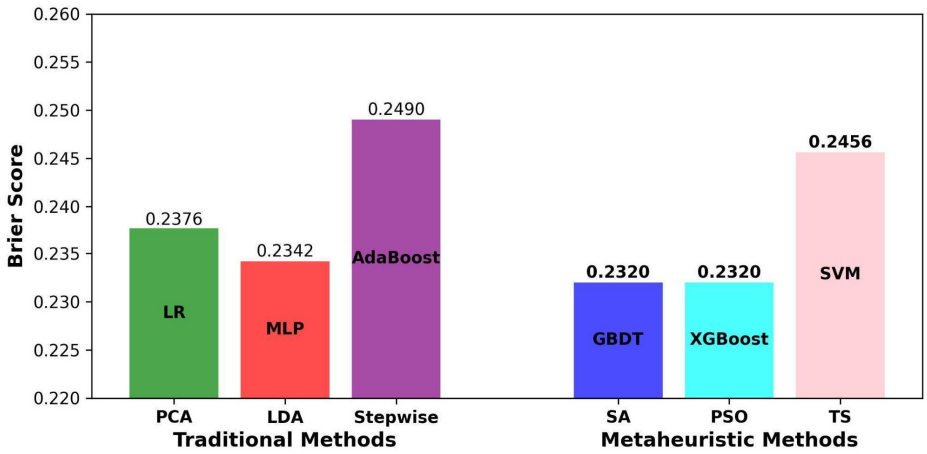

Fig. 4. Brier Score comparison of metaheuristic and traditional feature selection methods on the German dataset.

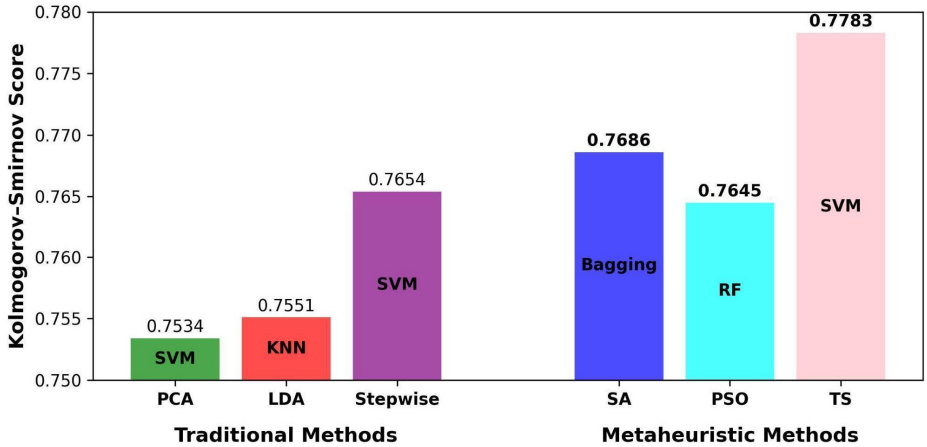

Fig. 5. Kolmogorov–Smirnov score comparison of metaheuristic and traditional feature selection methods on the Japanese dataset.

Overall, these results demonstrate that the metaheuristic-driven feature selection approach improves performance compared to both the baseline and other traditional feature selection methods. Specifically, the three methods, PSO, SA, and TS, achieve the best results across five evaluation metrics, with SA showing the most consistent performance across all three datasets. Moreover, the ensemble machine learning models, particularly when combining SA with RF, GBDT, or Bagging, exhibit superior performance compared to the others. Although this combination may not yield the best result on a single dataset compared with PSO or TS, it consistently improves performance relative to both the baseline and traditional methods.

IV. CONCLUSIONS

In conclusion, this study is positioned within the current context of recurring financial and debt crises, where large-scale and high-dimensional data make credit risk assessment increasingly complex. Our work demonstrates that integrating metaheuristic-driven feature selection provides a more effective means of filtering relevant information compared to traditional techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), or Stepwise Selection. Furthermore, combining these optimized features with machine learning models offers a robust and efficient framework for credit scoring.

By systematically removing irrelevant or redundant input data, the proposed pipeline not only enhances predictive accuracy but also improves model stability across diverse datasets. Among the tested approaches, the integration of Simulated Annealing (SA) with ensemble classifiers, particularly Random Forest (RF), Gradient Boosting Decision Trees (GBDT), and Bagging, consistently yielded the most reliable results, underscoring the practical value of the proposed methodology.

A comprehensive evaluation across multiple datasets and performance metrics further confirms the effectiveness and adaptability of metaheuristic-based feature selection in credit scoring applications. Future research may focus on exploring more advanced feature selection strategies, integrating them with transformer-based architectures, and refining data preprocessing techniques, such as outlier handling and class imbalance mitigation, to further enhance model robustness and generalization in financial prediction tasks.

REFERENCES

[1] J. C. Hull, *Risk Management and Financial Institutions*, 6th ed. Hoboken, NJ, USA: John Wiley & Sons, 2023.

[2] R. Sengupta and G. Bhardwaj, "Credit Scoring and Loan Default," *International Review of Finance*, vol. 15, no. 2, pp. 139–167, June 2015, https://doi.org/10.1111/irfi.12048.

[3] F. Louzada, A. Ara, and G. B. Fernandes, "Classification methods applied to credit scoring: Systematic review and overall comparison," *Surveys in Operations Research and Management Science*, vol. 21, no. 2, pp. 117–134, Dec. 2016, https://doi.org/10.1016/j.sorms.2016.10.001.

[4] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, vol. 91, June 2020, Art. no. 106263, https://doi.org/10.1016/j.asoc.2020.106263.

[5] Z. Song and H. Mo, "Transformative Impacts of Big Data Technologies on the Credit Reporting Industry: Drivers, Challenges, and Future Trajectories," *Journal of Frontier in Economic and Management Research*, vol. 1, no. 1, pp. 292–314, Sept. 2025, https://doi.org/10.63944/yae.JFEMR.

[6] W. A. Addy, A. O. Ajayi-Nifise, B. G. Bello, S. T. Tula, O. Odeyemi, and T. Falaiye, "AI in credit scoring: A comprehensive review of models and predictive analytics," *Global Journal of Engineering and Technology Advances*, vol. 18, no. 2, pp. 118–129, Feb. 2024, https://doi.org/10.30574/gjeta.2024.18.2.0029.

[7] A. A. H. Raji, A. H. F. Alabdoon, and A. Almagtome, "AI in Credit Scoring and Risk Assessment: Enhancing Lending Practices and Financial Inclusion," in *2024 International Conference on Knowledge Engineering and Communication Systems*, Chikkaballapur, India, 2024, pp. 1–7, https://doi.org/10.1109/ICKECS61492.2024.10616493.

[8] A. Ampountolas, T. Nyarko Nde, P. Date, and C. Constantinescu, "A Machine Learning Approach for Micro-Credit Scoring," *Risks*, vol. 9, no. 3, Mar. 2021, Art. no. 50, https://doi.org/10.3390/risks9030050.

[9] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1192, Mar. 2022, https://doi.org/10.1016/j.ejor.2021.06.053.

[10] O. A. Bello, "Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis," *International Journal of Management Technology*, vol. 10, no. 1, pp. 109–133, Dec. 2023, https://doi.org/10.37745/ijmt.2013/vol10n1109133.

[11] M. Ala'raj, M. F. Abbod, M. Majdalawieh, and L. Jum'a, "A deep learning model for behavioural credit scoring in banks," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5839–5866, Apr. 2022, https://doi.org/10.1007/s00521-021-06695-z.

[12] F. Shen, X. Zhao, G. Kou, and F. E. Alsaadi, "A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique," *Applied Soft Computing*, vol. 98, Jan. 2021, Art. no. 106852, https://doi.org/10.1016/j.asoc.2020.106852.

[13] C. X. Pham, H. N. Trinh, and L. Q. Tran, "A Robust Approach to Credit Scoring with Deep Learning and Embedded Methods," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 29284–29291, Dec. 2025, https://doi.org/10.48084/etasr.12649.

[14] V. Moscato, A. Picariello, and G. Sperlí, "A benchmark of machine learning approaches for credit score prediction," *Expert Systems with Applications*, vol. 165, Mar. 2021, Art. no. 113986, https://doi.org/10.1016/j.eswa.2020.113986.

[15] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the Operational Research Society*, vol. 54, no. 6, pp. 627–635, June 2003, https://doi.org/10.1057/palgrave.jors.2601545.

[16] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, Nov. 2015, https://doi.org/10.1016/j.ejor.2015.05.030.

[17] Y. Hayashi, "Emerging Trends in Deep Learning for Credit Scoring: A Review," *Electronics*, vol. 11, no. 19, Oct. 2022, Art. no. 3181, https://doi.org/10.3390/electronics11193181.

[18] H. Demma Wube, S. Zekarias Esubalew, F. Fayiso Weldesellasie, and T. Girma Debelee, "Deep Learning and Machine Learning Techniques for Credit Scoring: A Review," in *Second Pan-African Conference on Artificial Intelligence*, Addis Ababa, Ethiopia, 2023, pp. 30–61, https://doi.org/10.1007/978-3-031-57639-3_2.

[19] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Computational Economics*, vol. 57, no. 1, pp. 203–216, Jan. 2021, https://doi.org/10.1007/s10614-020-10042-0.

[20] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin, "Explainable AI in Credit Risk Management." arXiv, Mar. 01, 2021, https://doi.org/10.48550/arXiv.2103.00949.

[21] G. Manikandan, S. Abirami, O. P. Jena, A. R. Tripathy, A. A. Elngar, and Z. Polkowski, "Feature Selection and Machine Learning Models for High-Dimensional Data: State-of-the-Art," in *Computational Intelligence and Healthcare Informatics*, Hoboken, NJ, USA: John Wiley & Sons, Ltd, 2021, pp. 43–63, https://doi.org/10.1002/9781119818717.ch3.

[22] Y. Wu and Z. Huang, "Feature selection considering feature relevance, redundancy and interactivity for neighbourhood decision systems," *Neurocomputing*, vol. 596, Sept. 2024, Art. no. 128092, https://doi.org/10.1016/j.neucom.2024.128092.

[23] J. Laborda and S. Ryoo, "Feature Selection in a Credit Scoring Model," *Mathematics*, vol. 9, no. 7, Apr. 2021, Art. no. 746, https://doi.org/10.3390/math9070746.

[24] Y. Zhou, M. S. Uddin, T. Habib, G. Chi, and K. Yuan, "Feature selection in credit risk modeling: an international evidence," *Economic Research-Ekonomska Istraživanja*, vol. 34, no. 1, pp. 3064–3091, Jan. 2021, https://doi.org/10.1080/1331677X.2020.1867213.

[25] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, https://doi.org/10.1016/j.compeleceng.2013.11.024.

[26] A. Biernacki, "Evaluating Filter, Wrapper, and Embedded Feature Selection Approaches for Encrypted Video Traffic Classification," *Electronics*, vol. 14, no. 18, Sept. 2025, Art. no. 3587, https://doi.org/10.3390/electronics14183587.

[27] A. Siham, S. Sara, and A. Abdellah, "Feature selection based on machine learning for credit scoring : An evaluation of filter and embedded methods," in *2021 International Conference on INnovations in Intelligent SysTems and Applications*, Kocaeli, Turkey, 2021, pp. 1–6, https://doi.org/10.1109/INISTA52262.2021.9548410.

[28] Á. Paz *et al.*, "Machine Learning and Metaheuristics Approach for Individual Credit Risk Assessment: A Systematic Literature Review," *Biomimetics*, vol. 10, no. 5, May 2025, Art. no. 326, https://doi.org/10.3390/biomimetics10050326.

[29] T. Dokeroglu, A. Deniz, and H. E. Kiziloz, "A comprehensive survey on recent metaheuristics for feature selection," *Neurocomputing*, vol. 494, pp. 269–296, July 2022, https://doi.org/10.1016/j.neucom.2022.04.083.

[30] V. G. Helder, T. P. Filomena, L. Ferreira, and G. Kirch, "Application of the VNS heuristic for feature selection in credit scoring problems," *Machine Learning with Applications*, vol. 9, Sept. 2022, Art. no. 100349, https://doi.org/10.1016/j.mlwa.2022.100349.

[31] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 2052–2064, Mar. 2014, https://doi.org/10.1016/j.eswa.2013.09.004.

[32] R. Estran, A. Souchaud, and D. Abitbol, "Using a genetic algorithm to optimize an expert credit rating model," *Expert Systems with Applications*, vol. 203, Oct. 2022, Art. no. 117506, https://doi.org/10.1016/j.eswa.2022.117506.

[33] D. Martens, T. Van Gestel, M. De Backer, R. Haesen, J. Vanthienen, and B. Baesens, "Credit rating prediction using Ant Colony Optimization," *Journal of the Operational Research Society*, vol. 61, no. 4, pp. 561–573, Apr. 2010, https://doi.org/10.1057/jors.2008.164.

[34] F. Atban, M. Y. Küçükkara, and C. Bayılmış, "Enhancing variational quantum classifier performance with meta-heuristic feature selection for credit card fraud detection," *The European Physical Journal Special Topics*, vol. 234, no. 15, pp. 3705–3718, Oct. 2025, https://doi.org/10.1140/epjs/s11734-025-01703-y.

[35] B. K. Verma and A. K. Yadav, "Advancing Software Vulnerability Scoring: A Statistical Approach with Machine Learning Techniques and GridSearchCV Parameter Tuning," *SN Computer Science*, vol. 5, no. 5, May 2024, Art. no. 595, https://doi.org/10.1007/s42979-024-02942-x.

[36] J. Inga and E. Sacoto-Cabrera, "Credit Default Risk Analysis Using Machine Learning Algorithms with Hyperparameter Optimization," in *VIII International Conference on Science, Technology and Innovation for Society*, Guayaquil, Ecuador, 2022, pp. 81–95, https://doi.org/10.1007/978-3-031-24327-1_8.

[37] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, Dec. 2013, https://doi.org/10.1109/TSMCB.2012.2227469.

[38] R. Meiri and J. Zahavi, "Using simulated annealing to optimize the feature selection problem in marketing applications," *European Journal of Operational Research*, vol. 171, no. 3, pp. 842–858, June 2006, https://doi.org/10.1016/j.ejor.2004.09.010.

[39] H. Zhang and G. Sun, "Feature selection using tabu search method," *Pattern Recognition*, vol. 35, no. 3, pp. 701–711, Mar. 2002, https://doi.org/10.1016/S0031-3203(01)00046-2.

[40] R. Quinlan, "Statlog (Australian Credit Approval)." UCI Machine Learning Repository, 1987, https://doi.org/10.24432/C59012.

[41] H. Hofmann, "Statlog (German Credit Data)." UCI Machine Learning Repository, 1994, https://doi.org/10.24432/C5NC77.

[42] C. Sano, "Japanese Credit Screening." UCI Machine Learning Repository, 1992, https://doi.org/10.24432/C5259N.