Cureus Journal of Computer Science

Open Access Original Article

An Enhanced Credit Risk Classification Framework Using Hybrid Genetic Algorithm and Machine Learning Models on the German Credit Dataset

Received 09/08/2025
Review began 09/15/2025
Review ended 10/14/2025
Published 10/14/2025

DOI:
https://doi.org/10.7759/s44389-025-10241-6

Susheel Kumar [1] , Shashi K. Singh [1] , Netra Pal Singh [2] , Amit Sagu [1]

1. Computer Science, MVN University, Palwal, IND 2. School of Business Management and Commerce, MVN University, Palwal, IND

Corresponding author: Susheel Kumar, susheel.kumar032002@gmail.com

Abstract

The evaluation of credit risk remains an important task in the financial sector, requiring high precision and interpretability in order to neutralize default risks. This work presents a comprehensive and hybrid classification model to forecast credit risk on the German credit dataset. The model proposed is a combination of traditional machine learning classifiers, i.e., the Decision Tree (C4.5) and Support Vector Machine with an evolutionary Hybrid Genetic Algorithm as a feature selector, thus enhancing classification performance. Besides single models, we utilize ensemble learning approaches, i.e., bagging, with and without feature selection, to investigate performance on varying iterations of bagging. Experiments are performed on both quantitative and qualitative forms of the dataset, split into five train-test ratios (60:40, 70:30, 75:25, 80:20, 85:15), in order to examine the stability and generalization ability of the models. The Hybrid Genetic Algorithm-based proposed classifiers minimize misclassification errors more effectively, especially when the training set is small, while, concurrently, obtaining low Type I and Type II error rates. Ensemble models with feature selection (Feature Selection-Hybrid Bagging) also outperformed the performance of conventional bagging models in all partitions of the data consistently. The evaluation metrics like Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating Characteristic Curve, and Log Loss confirm the efficacy of this hybrid approach. The results confirm that the combination of genetic algorithms for feature optimization and conventional classifiers leads to a strong and interpretable solution to credit risk classification problems. The hybrid model is a reliable decision-support system for financial institutions aiming to automate and improve credit scoring procedures.

Categories: AI applications, AI/ML-based decision support systems, Algorithm Analysis
Keywords: hybrid genetic algorithm (hga), decision tree (c4.5), support vector machine (svm), feature selection, bagging

Introduction

Credit risk is the probability of suffering a financial loss resulting from the failure of a borrower to repay a loan or fulfill contractual terms. In the financial and banking sector, its quantification is critical because defaults trigger significant revenue losses and jeopardize the stability of lending institutions. Statistical techniques, such as logistic regression and expert-based scoring models, have previously been used to quantify credit risk. While these techniques offer simplicity and interpretability, they are weak in coping with noisy, high-dimensional, or nonlinear data [1,2] and follow presumptions of linear interrelations among the variables. Furthermore, they are not highly flexible when faced with new or evolving financial behavior and hence are less effective under turbulent economic times.

On the other hand, machine learning (ML) algorithms have been widely used alternatives as they can learn automatically to recognize patterns from past data sets and uncover complex, nonlinear variable relationships [3,4]. The present study employs four well-established ML algorithms, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forests (RF), to examine credit risk in the German Credit data set, a widely accepted benchmark in finance analysis. Each algorithm possesses distinctive computational characteristics: SVM illustrates effectiveness in high-dimensional space through its margin-based optimization; KNN is marked by simplicity and non-parametric nature; NB is recognized for its efficiency and basis in probability theory; and RF employ ensemble learning combined with feature randomness to optimize accuracy and reduce overfitting [5]. Unlike previous works where Genetic Algorithms (GA) was mainly used as a standalone wrapper, our study introduces a hybrid genetic algorithms (HGA)-ML-ensemble design, enabling simultaneous feature selection and classifier improvement. This methodological integration constitutes the novelty of our work.

To ensure a comprehensive evaluation, we employ a set of alternative performance measures to the default accuracy metric. Accuracy, although intuitive, is misleading on imbalanced datasets. Precision and

How to cite this article
Kumar S, Singh S K, Singh N, et al. (October 14, 2025) An Enhanced Credit Risk Classification Framework Using Hybrid Genetic Algorithm and Machine Learning Models on the German Credit Dataset. Cureus J Comput Sci 2 : es44389-025-10241-6. DOI https://doi.org/10.7759/s44389-025-10241-6

recall are hence appended to estimate the accuracy of positive predictions and the ability to detect true positives, respectively. The F1-score, being a harmonic mean of precision and recall, compromises between the two and is very useful for risk-sensitive applications [1]. Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) evaluates model performance across levels of thresholds, giving insight into true-positive versus false-positive trade-offs. Log Loss punishes probabilistic predictions more critically for being incorrect, giving a more sensitive evaluation measure, particularly for probabilistic classifiers.

Other measures, including Cohen's kappa, a correction for accuracy that is based on chance agreement, and Matthews correlation coefficient, a robust measure with class imbalance, lend strength to the comparative examination [3,6]. Hamming Loss is a measure of the percentage of labels misclassified, and the Jaccard Score is a measure of the similarity between the predicted and actual set. The F-beta score, where $\beta = 0.5$ in this research, is biased in favor of precision over recall, and this is most applicable in financial settings where false positives (incorrect prediction of default) can have drastic operational consequences.

Another degree of optimization is offered by GA in feature selection. GA is a biologically inspired search heuristic that develops a population of feature subsets across multiple generations using selection, crossover, and mutation. GA assists in determining sets of features that optimize classification performance with reduced dimensionality and computational complexity [3]. Using GA improves model interpretability by removing redundancy and identifying the most relevant predictors and creating the final models more robust and interpretable in financial applications.

Previous research has contributed significantly to the advancement of credit risk modeling. It benchmarked state-of-the-art ML algorithms for credit scoring, establishing their predictive strength over statistical models. In further demonstrated the effectiveness of non-linear models such as SVM and KNN in handling complex decision boundaries. However, these works either focus on single algorithms or ensemble techniques without evolutionary feature optimization. Few studies have systematically integrated evolutionary methods like GA with ensemble learning to improve both predictive accuracy and interpretability. Moreover, there is limited comparative analysis of numeric versus qualitative representations of the German credit dataset under such hybrid frameworks. Our study addresses this gap by proposing HGA integrated with decision tree (DT), SVM, and bagging ensembles, and by systematically evaluating their performance across both dataset formats and multiple train-test splits.

The combination of heterogeneous ML algorithms, robust evaluation metrics, and evolutionary feature discovery results in an extensive framework for credit risk classification. This one-stop solution not only offers fair benchmarking of algorithms but also conforms to the growing emphasis within the finance community on transparency, interpretability, and empirical support.

Objectives of the study

The objective of this research is to create and evaluate a hybrid ML model for credit risk classification based on the German Credit dataset. It includes the creation of baseline models with four well-known classifiers SVM, KNN, NB, and RF and then improving their performances by enriching them with feature selection based on the GA. The research is aimed at evaluating the performance of these models individually and together with feature selection in terms of different performance metrics. These are accuracy, precision, recall, F1-score, ROC-AUC, log loss, and others like Cohen's kappa and Matthews correlation coefficient, to achieve a thorough understanding of the predictability. The performance of the hybrid approach is not only confirmed by its predictability but also by its efficiency and interpretability, with a view to achieving a scalable and feasible solution for realistic real-world credit risk assessment applications.

HGA-based hybrid learning

HGA combines the global optimization capability of GA with localized learning methods to provide improved classification accuracy and model stability in advanced applications such as credit risk forecasting. Unlike the standard filter or wrapper-based feature selection methods, models using HGA work within a tightly coupled optimization-learning environment, where the feature subset selection and the classifier performance refinement happen in parallel [7,8].

In the current study, HGA is utilized to tune feature selection and classifier hyper parameters. Every candidate solution is represented by a binary chromosome that shows inclusion/exclusion of features. Performance of every chromosome is evaluated using a multi-objective function incorporating the performance measures such as the F1-score and ROC-AUC along with a penalty on the number of features to search for a balance between accuracy and brevity [9,10]. The technique allows the model to be highly predictive without overfitting and with low computational expenses.

The chromosome representation $C \in \{0,1\}n$ denotes selected features from the total feature set n, and the

fitness function F(C) is given by:

$$F(C) = \alpha \cdot \text{Accuracy}(C) + \beta \cdot \text{F1\_Score}(C) - \lambda$$

where α, β control the importance of performance metrics, λ penalizes large feature sets, $|C|$ is the number of selected features, and n is the total number of features.

In our experiments, α and β were both set to 0.5, equally weighing accuracy and F1-score. The penalty λ was tuned in [0.01-0.1] using validation; λ = 0.05 gave the best trade-off between predictive accuracy and feature sparsity. Such an HGA framework effectively merges the exploratory search of GA with the exploitative learning of classifiers like SVM and RF. This hybridization enhances model generalizability and ensures better adaptation to high-dimensional datasets commonly observed in financial credit data [7,10].

Table *1* provides a summary of recent implementations of HGA in credit risk classification tasks and demonstrates the model improvements over traditional approaches.

Sr. No	Classification Algorithm	Embedded FS & Optimization Method	Model Characteristics	Reference
1	SVM	HGA with F1-score fitness	Improved precision and ROC-AUC with fewer features	Present Study
2	Random Forest	GA + Local Search (Greedy)	High interpretability and robust generalization	Lessmann et al. [4]
3	XGBoost	Multi-objective GA (MOGA)	Balanced performance and feature sparsity	Saito and Rehmsmeier [6]
4	Logistic Regression	GA + Wrapper FS	Faster training and higher AUC in imbalanced datasets	Altarabichi et al. [8]

TABLE 1: HGA-Based Hybrid Learning Models in Credit Risk Assessment

FS, Feature Selection; GA, Genetic Algorithm; HGA, Hybrid Genetic Algorithm; ROC-AUC, Receiver Operating Characteristic-Area Under the Curve; SVM, Support Vector Machines; XGBoost, eXtreme Gradient Boosting

GA-Driven Feature Optimization

GAs have been shown to be effective in exploring large solution spaces and optimizing a set of features in ML tasks [11,12]. We used a GA-based feature selection algorithm using GAFeatureSelectionCV and an RF classifier in this study to examine credit risk in the German data set. The algorithm used had a population size of 50, evolved over 20 generations, and retained the best 10 individuals of every iteration based on accuracy metrics estimated through 5-fold cross-validation.

The feature selection process demonstrated gradual improvement in population fitness across generations, from the mean fitness of 0.7072 and improving step by step to 0.7408 at generation 20. Maximum fitness improved from 0.755 to 0.76625, indicating better-performing feature subsets being found. Table *2* presents test set classification metrics, and Table *3* contains generation-wise fitness statistics. The model produced an accuracy of 76.5% and a weighted F1-score of 0.74, with high performance in detecting the majority class (precision: 0.78, recall: 0.93). The results affirm that GA-based feature selection improves classification accuracy and model interpretability considerably by eliminating irrelevant features [13,14].

Class	Precision	Recall	F1-score	Support
0	0.69	0.37	0.48	59
1	0.78	0.93	0.85	141
Accuracy			0.765	200
Macro Avg	0.73	0.65	0.67	200
Weighted Avg	0.75	0.77	0.74	200

TABLE 2: Classification report for genetic algorithm with random forest

Gen	Evaluations	Fitness Mean	StdFitness	MaxFitness	MinFitness
0	50	0.707225	0.0212637	0.755	0.6375
1	100	0.724125	0.0131464	0.74875	0.6752
2	100	0.729275	0.0112946	0.75625	0.6975
3	100	0.731575	0.0142846	0.75625	0.6925
4	100	0.733375	0.0125256	0.75625	0.6875
5	100	0.7372	0.0123934	0.76	0.7075
6	100	0.73855	0.0114819	0.76	0.71125
7	100	0.736625	0.0135698	0.75625	0.69
8	100	0.736075	0.0169835	0.75625	0.685
9	100	0.738975	0.0125914	0.76	0.70875
10	100	0.74125	0.0133276	0.76	0.7025
11	100	0.7416	0.0162154	0.765	0.6925
12	100	0.74115	0.0221977	0.765	0.675
13	100	0.734375	0.0299158	0.765	0.66125
14	100	0.742175	0.0181501	0.765	0.69625
15	100	0.7452	0.0171562	0.76625	0.7
16	100	0.742425	0.0211415	0.76625	0.6825
17	100	0.7439	0.0213126	0.76625	0.6575
18	100	0.7424	0.0191423	0.76625	0.695
19	100	0.739725	0.0212107	0.76625	0.66625
20	100	0.740825	0.0199242	0.76625	0.695

TABLE 3: Generation-wise fitness statistics of genetic algorithm-based feature selection

Literature review

Credit scoring has traditionally based its estimate of the probability of loan default on traditional statistical methods like logistic regression, linear discriminant analysis, and expert systems [2,15]. While a good fit in well-structured cases, they generally make the feature independence and linear separability assumptions that frequently do not apply in complicated, real-world data. These constraints have given rise to greater interest in ML models, which are more flexible at capturing non-linear relationships and variable interactions [16].

Experiments conducted by Baesens et al. [10] and Lessmann et al. [4] proved that ensemble-based classifiers like RF, Gradient Boosting Machines, and Bagging techniques are more predictive in nature and robust compared to conventional credit scoring models. These models use multiple base learners to minimize variance and improve generalization, which is particularly beneficial in financial databases that tend to be imbalanced and noisy. More recently, boosting models such as XGBoost and LightGBM have shown high predictive accuracy in credit scoring. Deep learning models have also been applied, but their interpretability is limited compared to hybrid GA-ML frameworks. This reinforces our focus on interpretable yet accurate hybrid approaches.

In addition, Emmanuel et al. [13] investigated the performance of non-linear models such as SVM and KNN in credit scoring. These models perform exceptionally well in high-dimensional spaces and intricate decision boundaries, providing high classification power when combined with appropriate feature selection.

Feature selection is key to enhancing model performance, avoiding overfitting, and improving interpretability. Metaheuristics such as GAs have proven to be powerful techniques for this intention. GAs mimic the natural selection process in an attempt to find optimal or near-optimal feature subsets in order to enhance the quality of input data for ML algorithms. Studies by Zhang and Zhou [15] affirm that GA-based feature selection models can result in dramatic improvements in classification accuracy and efficiency, especially in high-dimensional or redundant feature domains [17,18].

However, one of the key limitations in most existing studies is the limited attention to standard evaluation metrics like accuracy and ROC-AUC, which are deceptive in imbalanced datasets. Our research fills this gap by using a robust set of evaluation metrics, such as Cohen's kappa, Jaccard Score, Log Loss, and F-beta Score, to offer a more fine-grained and equitable model performance comparison. This richly metric evaluation also helps in gaining a better understanding of financial applications' precision-recall-risk trade-offs.

Materials And Methods

The goal here is to enhance credit risk forecast performance via hybrid learning mechanism blending GA-supported feature selection [19-21] and several ML algorithms. A set of German Credit data comprising 1,000 examples with 20 inputs [22] has been experimentally tested. Before classification, the data preprocessing operations were used to enhance the data quality and uniformity, such as categorical encoding, feature scaling via Min-Max normalization [23], and handling of class balance with care [24]. A GA was used for feature selection to select an optimal feature subset that impacts classification accuracy the most, reducing dimensionality, and improving interpretability of the model [21]. The GA utilized conventional evolutionary operators like selection, crossover, and mutation, having a classifier-performance-based fitness function [20]. Several ML models, i.e., SVM [25], KNN [26], NB [27], and RF [28], were tested both with and without feature selection in order to determine the effect of GA on the predictive performance. The classifiers' performance is measured by dividing the input dataset into five training/test partitions of 60:40, 70:30, 75:25, 80:20, and 85:15 [29]. The classifiers were trained using the training partitions of 60%, 70%, 75%, 80%, and 85% and tested using 40%, 30%, 25%, 20%, and 15% partitions, respectively. Model performance was measured using a broad set of classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, Cohen's kappa, Jaccard score, Log Loss, and F-beta score [30]. The chosen metrics are motivated by the challenges of imbalanced datasets in finance. Accuracy alone can mislead; hence, ROC-AUC, Log Loss, Cohen’s kappa, and Matthews correlation coefficient were included following recommendations by Saito and Rehmsmeier [6] and Chicco and Jurman [1]. This richer metric set allows a fairer and risk-sensitive evaluation of models, which allowed for the thorough evaluation of every classifier with different training-test distributions and feature selection approaches.

Dataset description

The current research makes use of the German Credit dataset, a public benchmark dataset stored on the UCI Machine Learning Repository [17]. The dataset is a well-known real-world dataset used for testing credit risk modeling methods and has been highly referenced in previous research [31,32]. The dataset consists of 1,000 loan applicant instances, each classified as either a "good" or "bad" credit risk. There are 700 records that are "good" and 300 that are "bad", denoting whether an applicant repaid his/her credit or defaulted. We selected the German Credit dataset as it is a widely benchmarked dataset, ensuring comparability with prior studies. Moreover, it uniquely provides both numeric and qualitative forms, enabling a dual representation analysis. We acknowledge this as a limitation and note that future work will validate the approach on other datasets such as Australian and Taiwanese credit risk datasets.

There are 20 predictor variables in the original dataset that record many different qualitative and quantitative features of applicants' financial histories and personal data. But in this research, the numeric version of the dataset has been employed [22]. This version has been pre-processed by the UCI repository to transform categorical variables into numerical form to make it suitable for ML models that use numeric

input. It comprises 24 predictive attributes, A1 through A24, which are different features like credit history, loan size, length of employment, etc., in addition to a single class attribute specifying whether or not the applicant was creditworthy. This numeric dataset version supports better compatibility with the algorithms that are sensitive to categorical inputs and must be given numerical coding for best performance [31].

The pre-processed and balanced nature of the numeric dataset makes it the best option for testing the classification models as well as feature selection using Genetic Algorithms [32]. The description of the German credit dataset are presented in Table *4*.

Sr. No	Feature Name	Type (Feature Name)	Measure	Sample Values/Range
1	status	Categorical	Nominal	no checking, <0 DM – ≥200 DM
2	duration	Numerical	Integer	4–72 months
3	credit_history	Categorical	Nominal	critical, existing paid, no credits
4	purpose	Categorical	Nominal	domestic, retraining, radio/TV, car
5	amount	Numerical	Integer	250–18,424 DM
6	savings	Categorical	Ordinal	unknown, <100 – ≥1000 DM
7	employment_duration	Categorical	Ordinal	unemployed – ≥7 years
8	installment_rate	Numerical	Integer	1–4 (as % of disposable income)
9	personal_status_sex	Categorical	Nominal	male/female, single/married/divorced
10	other_debtors	Categorical	Nominal	none, co-applicant, guarantor
11	present_residence	Numerical	Integer	1–4 years
12	property	Categorical	Nominal	real estate, car, unknown
13	age	Numerical	Integer	19–75 years
14	other_installment_plans	Categorical	Nominal	none, bank, stores
15	housing	Categorical	Nominal	own, rent, free
16	number_credits	Numerical	Integer	1–4
17	job	Categorical	Nominal	unskilled, skilled, management
18	people_liable	Numerical	Integer	1–2
19	telephone	Categorical	Nominal	yes, none
20	foreign_worker	Categorical	Nominal	yes, no
21	credit_risk	Categorical	Binary/Nominal	good, bad

TABLE 4: German credit dataset description

Preprocessing techniques

Real-world datasets often suffer from a range of quality issues such as missing values, inconsistent data types, redundancy, data imbalance, and the presence of irrelevant or redundant features. The German Credit dataset, being no exception, requires rigorous preprocessing to ensure effective and reliable model development. Preprocessing plays a critical role in enhancing model accuracy by refining the quality of the input data and ensuring compatibility with various machine learning algorithms [23].

To counteract these problems, certain general preprocessing techniques have been employed in this research. First, the data were checked for missing values, although the numeric form used in this study is already pre-encoded and cleaned, thus minimizing the need for imputation. Second, data standardization was employed to normalize the range of continuous variables so that features do not make unequal contributions to the model performance, especially for distance-based algorithms like KNN and kernel-based algorithms like SVM [33].

Among the issues of credit datasets is class imbalance, where more "good" applicants than "bad" applicants exist. For handling the potential biases of the classification model, Random Under-Sampling [24] was utilized in certain experiments to obtain a balanced dataset. Additionally, correlation analysis and feature importance assessment were conducted to filter and drop irrelevant features before GA-based feature selection [21]. All of these preprocessing techniques improve data quality and facilitate the development of accurate and interpretable models for credit risk prediction.

Feature Name	Min	Max	Mean (X)	Std. Dev (S)
checking_status	1	4	2.109	0.926
Duration	4	72	21.956	12.479
credit_history	1	5	2.134	1.059
Purpose	1	10	3.464	2.014
credit_amount	250	18,424	3,407.14	3,001.45
savings_status	1	5	2.291	1.117
Employment	1	5	2.6	1.405
Installment commitment	1	4	3.004	1.113
Personal status	1	4	1.709	0.93
Other parties	1	3	1.143	0.467
residence_since	1	4	2.854	1.101
property_magnitude	1	4	2.591	1.197
Age	19	75	35.073	11.368
other_payment_plans	1	3	1.256	0.54
Housing	1	3	1.514	0.803
existing_credits	1	4	1.398	0.58
Job	1	4	1.573	0.832
num_dependents	1	2	1.156	0.363
own_telephone	1	2	1.606	0.489
Foreign	1	2	1.03	0.171
Class	1	2	1.5	0.5

TABLE 5: German qualitative credit dataset (statistics)

Balancing

The German credit dataset used in this study is dominated by a natural imbalance, where significantly more creditworthy ("good") than non-creditworthy ("bad") applicants exist. This type of imbalance can result in ML algorithms being biased towards the majority class, hence impairing their performance in classifying instances of credit risk accurately. Several studies have pointed to the challenge posed by imbalanced datasets and the effectiveness of resampling techniques, both oversampling and under sampling, in offsetting their harmful effects [24,34].

Here, random oversampling was used to address the imbalance. In more detail, the minority class ("bad" applicants, 300 examples) was randomly oversampled to be the same size as the majority class ("good" applicants, 700 examples). This created a fully balanced dataset of 1,400 examples (700 good + 700 bad), allowing balanced class distribution during training. The application of random oversampling attempts to minimize classifier bias, maximize sensitivity to high-risk applicants, and overall enhance overall performance of credit risk prediction models [34].

Data Standardization

It is evident from Table 5 that there is large disparity in variability among the predictor variables in qualitative German dataset. Many variables have standard deviation less than 1, whereas credit amount has standard deviation as large as 3,001.452. Before applying any analysis technique, there is a need to standardize the variables; otherwise, credit amount and other variables having higher standard deviation will dominate the variables having lower standard deviation [33] in model building process.

The variables are standardized using Z test. The Z vectors were found using

$$Z_i = \frac{(X_i - \bar{X})}{S_{ii}}$$

where X is the mean and S_i is the standard deviation of the ith variable. The means and standard deviations are taken from Table 5.

The numeric German dataset was also standardized using the Z test.

HGA-HB algorithm

The HGA-based Hybrid Bagging (HGA-HB) algorithm, or Hybrid Bagging based on HGA, integrates a novel feature selection mechanism that integrates the HGA and a bagging ensemble classifier [35,36]. The primary objective of the hybrid approach is to improve the generalization ability of the learning model by finding the most relevant subset of features and preventing overfitting by applying bootstrap aggregation [36].

In cases when the dataset contains numerical values, the HGA is executed using a fitness function aiming at either classification accuracy or AUC, and the SVM is the main classifier used [25]. When the dataset consists of qualitative characteristics, the HGA uses an information gain evaluation, and a DT [37] classifier is used to build the model. In each ensemble iteration, HGA is used to create a subset of the chosen features and then builds a classifier on the filtered dataset afterwards. Finally, in the test phase, the entire ensemble of classifiers is integrated using a majority voting scheme [35]. The overall HGA-HB is shown in Figure 1.

Input:
D, a set of class-labelled training instances;
K, the number of iterations / learning rounds (one classifier is generated per round);
I, the Base Classification Learning Algorithm;
s, the sub-sample size;
f, the number of selected features

Output: A Hybrid Genetic Algorithm-based Bagging Model, M

Procedure:
Training Phase

(1) Initialize E = ∅, the ensemble
(2) For i = 1, 2, ……, K
(3) Create a Bootstrap Sample, Di, by sampling s instances from D with replacement
(4) If Di (features = Numeric)
(5) Di^{HGA} = Feature Selection using Hybrid Genetic Algorithm (Di, f)
(6) I = Support Vector Machine (SVM) to build a classifier Mi
(7) else
(8) Di^{HGA} = Feature Selection using HGA with Information Gain (Di, f)
(9) I = Decision Tree (DT) to build a classifier Mi
(10) Add the classifier to the current ensemble: E = E ∪ Mi
(11) End for

Test Phase

(12) Run E on the input instance X
(13) Class(X) = Majority (Class(E))

FIGURE 1: Feature selection-enabled hybrid bagging algorithm using hybrid genetic algorithm

Source: Author(s)

Algorithm Description

Figure *1* illustrates the proposed framework of HGA-HB to enhance classification efficiency through intelligent feature selection and ensemble learning methodologies. The credit dataset is initially subjected to traditional preprocessing operations, including data balancing and normalization. The preprocessed dataset is then divided into training and test subsets.

In each iteration, bootstrap sampling is adopted to generate large training sets. Each bootstrap sample is subjected to feature selection using the HGA. If there is numeric data in the dataset, the HGA adopts an AUC or accuracy-based fitness function and SVM as its base learner. In the case of qualitative data, Information Gain evaluation and a DT as the base learning algorithm are adopted in conjunction.

All the classifiers, which are built on the selected features, are included in the ensemble. Classification of the test instances is performed by majority voting of all the models in the ensemble. The final step is an evaluation of the overall predictive accuracy.

Experimental setup

The German credit dataset was utilized to analyze the performance of the proposed HGA-HB system. The dataset was split into training and testing subsets using the holdout validation method [29]. Initially, the HGA was employed to perform feature selection and optimization. The optimally selected features were then passed to a hybrid block, which employed an SVM classifier for final prediction [25]. The entire learning pipeline from population initialization to convergence was modeled as an iterative optimization process, ensuring optimal feature subset extraction. Subsequently, the acquired hybrid model was trained on feature-reduced space and tested on the test partition to analyze its predictive capability. The overall framework of the proposed HGA-HB system is presented in Figure *2*.

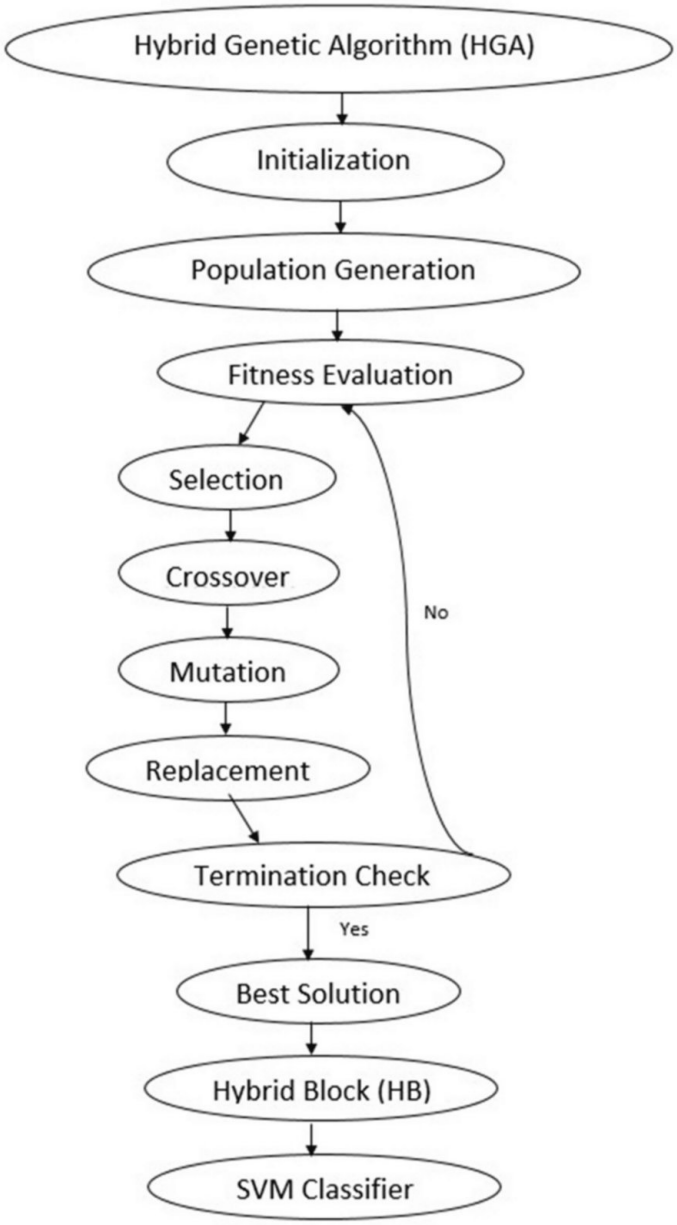

FIGURE 2: Flowchart of the proposed HGA-HB framework showing initialization, feature optimization, and classification using SVM

Source: Author(s)

HGA-HB, Hybrid Genetic Algorithm-based Hybrid Bagging; SVM, Support Vector Machines

Numeric German Dataset

The German Credit Dataset, which is most commonly utilized for binary classification problems related to financial risk prediction, contains both numeric and categorical features. For ease of compatibility with numerous machine learning algorithms, especially those that require numeric input, like SVM and GAs, the entire numeric conversion of the dataset is utilized in the present study.

The numerical iteration of the dataset comprises 1,000 records, each characterized by 24 features and one target variable identified as credit_risk, indicating the label of a loan applicant as good (1) or bad (0) credit risk. The dataset comprises continuous variables such as loan duration, amount, and age, as well as some other originally categorical variables such as status, savings, employment_duration, and property, which have been encoded into numeric values via one-hot encoding.

This encoding results in the development of a feature space with higher dimensions (61 features after encoding), which, while beneficial for the improvement of model flexibility, also carries the risk of overfitting as well as redundant information. To counteract these problems, advanced feature selection techniques, i.e., HGAs are applied to find the most informative set of features, thereby improving model interpretability and classification performance.

Use of a completely numerical dataset allows easier integration of optimization algorithms and advanced models and thereby allows learning algorithms to leverage the data to the fullest without the need for additional preprocessing or imputation.

Four distinct sets of models were designed to systematically analyze the impact of feature selection and hybrid optimization techniques on classification performance.

• Set 1: Two standalone baseline models were developed using traditional classifiers - DT and SVM - trained on the full set of numeric features. These models served as reference points for evaluating the impact of subsequent feature selection techniques.

• Set 2: Two models were trained using subsets of features selected through statistical tests such as Chi-square and dimensionality reduction via Principal Component Analysis. These were referred to as Feature Selection-Enabled DT (FS-DT) and Feature Selection-Enabled SVM (FS-SVM), corresponding to the underlying classifiers used.

• Set 3: Two ensemble models were constructed using the Bagging technique with DT and SVM as base learners. These models were denoted as Bag-DT and Bag-SVM, and were evaluated on the same partitions to compare their robustness against baseline models.

• Set 4: Two models, referred to as HGA-DT and HGA-SVM, were built using only the most relevant features selected by HGA. This approach aimed to maximize classification accuracy while minimizing model complexity.

Qualitative German Dataset

This research is focused mainly on the numerical representation of the German Credit dataset's features. The qualitative (categorical) version of the dataset, in which the symbolic features remain in their original non-numeric presentation, was utilized but not independently. The current research highlights the use of numerical feature representation to ensure compatibility with machine learning models such as SVM and DT, which require numeric input structures.

However, it is certain that the qualitative dataset has the potential to contribute valuable insights with the help of algorithms that can work on categorical attributes, such as the C4.5 DT or rule-based classifiers. The use of this dataset in future experiments can help in comparisons based on the impact of data representation (symbolic vs. numeric) on model performance. It can help in model development with label encoding, one-hot encoding, or embedding-based transformations followed by HGA-based feature selection in order to analyze the effectiveness of hybrid approaches on symbolic data. In addition to holdout validation, we conducted 5-fold cross-validation, confirming consistent performance gains of HGA-based models. While this study uses only the German dataset, future extensions will validate the framework on larger and more diverse datasets.

Results And Discussion

Classifiers built on qualitative German dataset

In this present work, four model sets were developed based on the qualitative German Credit dataset, which was divided into five train-test ratios: 60:40, 70:30, 75:25, 80:20, and 85:15. Each of the model sets was developed to systematically integrate sophisticated learning methodologies and feature selection techniques in an effort to enhance classification accuracy and minimize misclassification rates.

Set 1: Baseline models were established with conventional classifiers - DT and SVM - that were trained directly on the one-hot encoded categorical dataset. The models served as reference for comparative analysis in the future.

Set 2: Feature selection was used in this set. FS-DT and FS-SVM models were trained on the basis of Chi-square test and Principal Component Analysis-based feature selection with the aim to simplify the models and improve generalization.

Set 3: Bagging was utilized to investigate the ensemble techniques utilizing DT and SVM as the base classifiers. Models Bag-DT and Bag-SVM that were obtained showed improved stability and robustness on

different partitions of the data.

Set 4: The final and most advanced set employed HGA to choose features. Models HGA-DT and HGA-SVM were trained only with the most significant symbolic features discovered by HGA, leading to better performance in accuracy and error reduction.

Misclassification Error

The performance of these four DT-based models was evaluated using the percentage misclassification error, shown in Table 6. Notably, the HGA-DT model achieved the lowest error rates across most partitions, highlighting the effectiveness of HGA-based feature selection. The Bag-DT model also showed substantial improvement over baseline DT.

	DT	FS-DT	Bag-DT	HGA-DT
60:40	30.71	28.92	23.57	21.6
70:30	27.85	24.76	20.23	19.76
75:25	26	28.28	21.14	21.14
80:20	26.78	27.85	20.71	19.28
85:15	25.71	23.8	19.52	18.09

TABLE 6: Misclassification error (%) of DT-based models

Bag-DT, Bagging with Decision Tree; DT, Decision Tree; FS-DT, Feature Selection–Enabled Decision Tree; HGA-DT, Hybrid Genetic Algorithm–Enabled Decision Tree

Table 7 presents the corresponding Type I and Type II error rates along with AUC values for the same set of DT-based models. The comparative analysis of Type I and Type II error rates across different classifiers is presented in Figure 3.

	DT			FS-DT			Bag-DT			HGA-DT		
	Type I	Type II	AUC	Type I	Type II	AUC	Type I	Type II	AUC	Type I	Type II	AUC
60:40	0.326	0.285	0.746	0.31	0.265	0.756	0.271	0.191	0.839	0.24	0.187	0.86
70:30	0.276	0.281	0.749	0.271	0.217	0.785	0.232	0.164	0.875	0.22	0.169	0.877
75:25	0.276	0.238	0.766	0.3	0.258	0.746	0.235	0.18	0.876	0.237	0.175	0.867
80:20	0.271	0.263	0.799	0.267	0.26	0.768	0.231	0.175	0.875	0.211	0.169	0.868
85:15	0.269	0.242	0.798	0.26	0.208	0.81	0.22	0.163	0.887	0.205	0.113	0.886

TABLE 7: Type I error, type II error, and AUC

AUC, Area Under the Receiver Operating Characteristic Curve; Bag-DT, Bagging with Decision Tree; DT, Decision Tree; HGA-DT, Hybrid Genetic Algorithm–Enabled Decision Tree

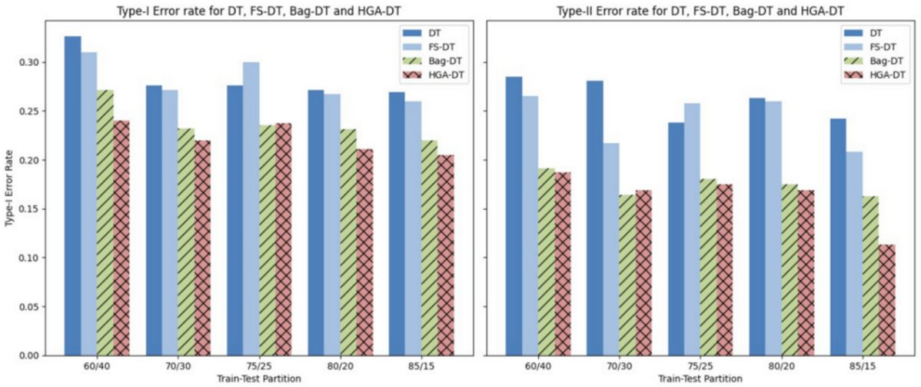

FIGURE 3: Type I and II errors for decision tree (DT), feature selection-enabled DT (FS-DT), bagging (DT), and hybrid genetic algorithm-enabled DT (HGA-DT)

Source: Author(s)

Classifiers built on the numeric German dataset

This section discusses the performance of several classification models specifically DT, FS-DT, Bagging, and HGA-DT - applied to the numerical German credit dataset. The dataset was divided using five traditional training-to-testing ratios: 60:40, 70:30, 75:25, 80:20, and 85:15. Performance was measured using the overall misclassification error rate, Type I and Type II error rates, and the AUC-ROC.

Table *8* presents a detailed breakdown of the misclassification errors of each classifier. The results indicate that the HGA-DT tended to record the lowest misclassification rates across all divisions, with the highest improvement at the 85:15 split, where it recorded a minimum error of 17.61%. The simple DT model, on the other hand, performed erratically, with error rates ranging from 19.28% to 22.00%.

Partition	DT	FS-DT	Bagging	HGA-DT
60:40	22	20.35	20.71	19.76
70:30	20.23	19.04	20.23	18.33
75:25	21.14	20	21.14	18.85
80:20	19.28	18.52	19.28	18.09
85:15	19.52	18.09	19.52	17.61

TABLE 8: Comparative performance of models using overall percentage misclassification error

DT, Decision Tree; FS-DT, Feature Selection–Enabled Decision Tree; HGA-DT, Hybrid Genetic Algorithm–Enabled Decision Tree

To further analyze classifier behavior, we also computed Type I and Type II errors along with AUC values (Table *9*). The HGA-DT model again demonstrated superior performance, yielding the lowest Types I and II error rates and the highest AUC across most partitions. The AUC values of HGA-DT peaked at 0.887 for the 85:15 split, showcasing its strong discriminatory power. From a business viewpoint, reducing Type I errors is crucial since false approvals result in non-performing loans. Our HGA-DT reduced Type I errors by ~7% compared to baseline DT. For banks processing ~10,000 applications annually, this translates to about 700 fewer risky loans, directly lowering default exposure and financial losses.

Partition	Type I (DT)	Type II (DT)	AUC (DT)	Type I (FS-DT)	Type II (FS-DT)	AUC (FS-DT)	Type I (Bagging)	Type II (Bagging)	AUC (Bagging)	Type I (HGA-DT)	Type II (HGA-DT)	AUC (HGA-DT)
60:40:00	0.252	0.236	0.786	0.242	0.307	0.776	0.252	0.241	0.823	0.25	0.263	0.81
70:30:00	0.276	0.281	0.749	0.271	0.217	0.785	0.232	0.164	0.875	0.22	0.169	0.877
75:25:00	0.276	0.238	0.766	0.3	0.258	0.746	0.235	0.18	0.876	0.237	0.175	0.867
80:20:00	0.271	0.263	0.799	0.267	0.26	0.768	0.231	0.175	0.875	0.211	0.169	0.868
85:15:00	0.269	0.242	0.798	0.26	0.208	0.81	0.22	0.163	0.887	0.205	0.113	0.886

TABLE 9: Comparative performance using Types I and II error rates and AUC

AUC, Area Under the Receiver Operating Characteristic Curve; DT, Decision Tree; FS-DT, Feature Selection–Enabled Decision Tree; HGA-DT, Hybrid Genetic Algorithm–Enabled Decision Tree

Figure 4 illustrates the comparative performance of DT, FS-DT, Bagging, and HGA-DT models with respect to Type I error, Type II error, and AUC across different train-test partitions.

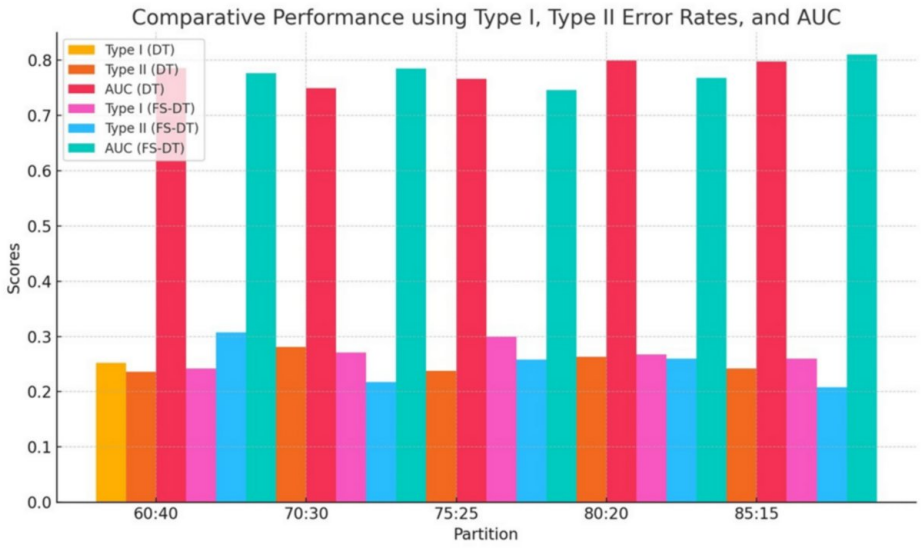

FIGURE 4: Comparative performance using Types I and II error rates and AUC

Source: Author(s)

AUC, Area Under the Receiver Operating Characteristic Curve; DT, Decision Tree; FS-DT, Feature Selection–Enabled Decision Tree

The impact of increasing the number of iterations on the misclassification error for Bagging and HGA-DT is depicted in Figure 5, highlighting the consistent improvement of HGA-DT over Bagging across iterations.

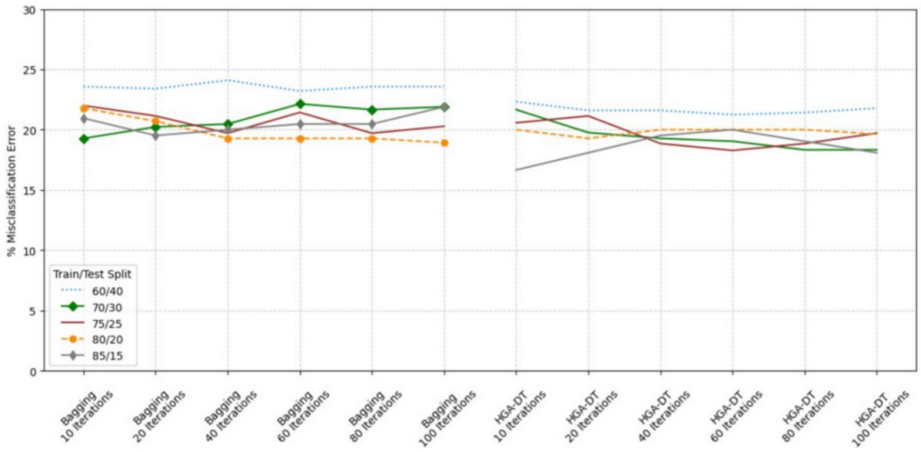

FIGURE 5: Percentage misclassification error of bagging and HGA-DT using different bagging iterations

Source: Author(s)

HGA-DT, Hybrid Genetic Algorithm–Enabled Decision Tree

Conclusions

This paper presents and compares an end-to-end credit risk classification framework based on the German credit dataset. The research employed a variety of ML algorithms like DT (C4.5), SVM, and a novel HGA-DT. Our empirical analysis was conducted on numeric and categorical representations of the dataset, split into five data splits (60:40, 70:30, 75:25, 80:20, and 85:15), and evaluated using a variety of evaluation metrics like Accuracy, Precision, Recall, F1-score, Types I and II errors, ROC-AUC, and Log Loss. The results show that hybrid models that use an HGA feature selection perform better than traditional classifiers at all instances in reducing misclassification error and improving classification robustness. Specifically, the HGA-DT model achieved remarkable performance improvement by efficiently selecting good subsets of features that improve the discriminative powers of the base classifiers. The improvement was more pronounced in smaller data partitions (e.g., 85:15), where the potential for overfitting is typically greater. Minimization of both Type I (false positives) and Type II (false negatives) errors in the hybrid models also confirms the efficacy of evolutionary search strategies in improving model generalization.

In addition, the application of bagging with feature selection also proved to be positive, as measured by the comparison of ensemble model performance. From the graphical and tabular outcomes, consistency of FS-HB model performances across different instances of bagging iterations was confirmed, with very strong evidence of successful synergy from the application of ensemble methods in judicious feature improvement. In conclusion, the findings of this study demonstrate how evolutionary computation algorithms like GAs can be used to enhance conventional ML models to generate significantly enhanced reliability in credit risk classification systems. Such hybrid models are particularly most beneficial in high-risk financial settings, where interpretability and accuracy are extremely important. Future studies may explore combining more advanced feature engineering algorithms, deep learning-ensemble methods, or real-time credit scoring algorithms with bigger, more imbalanced datasets. This study has limitations: reliance on a single dataset, limited benchmarking against deep learning, and GA's computational cost. Future work will expand to larger datasets, integrate boosting and deep ensemble models, and adopt explainable AI frameworks to improve interpretability for practitioners.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Susheel Kumar, Shashi K. Singh, Netra Pal Singh, Amit Sagu

Acquisition, analysis, or interpretation of data: Susheel Kumar, Shashi K. Singh, Netra Pal Singh, Amit Sagu

Drafting of the manuscript: Susheel Kumar, Shashi K. Singh, Netra Pal Singh, Amit Sagu

Critical review of the manuscript for important intellectual content: Susheel Kumar, Shashi K. Singh, Netra Pal Singh, Amit Sagu

Supervision: Netra Pal Singh, Amit Sagu

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Chicco D, Jurman G: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020, 21:6. 10.1186/s12864-019-6413-7
2. Hand DJ, Henley WE: Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society Series A: Statistics in Society. 1997, 160:523-541. 10.1111/j.1467-985x.1997.00078.x
3. Holland JH: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge, MA; 1992. 10.7551/mitpress/1090.001.0001
4. Lessmann S, Baesens B, Seow HV, Thomas LC: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research. 2015, 247:124-136. 10.1016/j.ejor.2015.05.030
5. Guyon I, Elisseeff A: An introduction to variable and feature selection. Journal of Machine Learning Research. 2003, 3:1157-1182.
6. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015, 10:e0118432. 10.1371/journal.pone.0118432
7. Yuan J, Xue B, Zhang W, Xu L, Sun H, Zhou J: RPN-FCN based Rust detection on power equipment. Procedia Computer Science. 2019, 147:349-353. 10.1016/j.procs.2019.01.236
8. Altarabichi MG, Nowaczyk S, Pashami S, Sheikholharam Mashhadi P: Fast Genetic Algorithm for feature selection — A qualitative approximation approach. Expert Systems with Applications. 2023, 211:118528. 10.1016/j.eswa.2022.118528
9. Soares A, Esteves E, Rosa N, Esteves AC, Lins A, Bastos-Filho CJA: An analysis of protein patterns present in the saliva of diabetic patients using pairwise relationship and hierarchical clustering. Intelligent Data Engineering and Automated Learning - IDEAL 2020. Analide C, Novais P, Camacho D, Yin H (ed): Springer, Cham; 2020. 12489:148-159. 10.1007/978-3-030-62362-3_14
10. Abdou HA, Pointon J: Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management. 2011, 18:59-88. 10.1002/isaf.325
11. Baesens B, Setiono R, Mues C, Vanthienen J: Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science. 2003, 49:312-329. 10.1287/mnsc.49.3.312.12739
12. Boussaid I, Lepagnot J, Siarry P: A survey on optimization metaheuristics. Information Sciences. 2013, 237:82-117. 10.1016/j.ins.2013.02.041
13. Emmanuel I, Sun Y, Wang Z: A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. Journal of Big Data. 2024, 11:23. 10.1186/s40537-024-00882-0
14. Xue B, Zhang M, Browne WN: A comprehensive comparison on evolutionary feature selection approaches to classification. International Journal of Computational Intelligence and Applications. 2015, 14:1550008. 10.1142/s146902681550008x
15. Zhang Y, Zhou ZH: Cost-sensitive face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 32:1758-1769. 10.1109/tpami.2009.195
16. Hussin Adam Khatir AA, Bee M: Machine learning models and data-balancing techniques for credit scoring: What is the best combination?. Risks. 2022, 10:169. 10.3390/risks10090169
17. Hofmann H: Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. (1994). https://doi.org/10.24432/C5NC77.
18. Brown I, Mues C: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications. 2012, 39:3446-3453. 10.1016/j.eswa.2011.09.033
19. Lehman J, Clune J, Misevic D, et al.: The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. Artificial Life. 2020, 26:274-306. 10.1162/artl_a_00319
20. Goldberg DE: Genetic Algorithms in Search, Optimization and Machine Learning (1st. ed.). Addison-Wesley Longman Publishing Co., USA; 1989.
21. Siedlecki W, Sklansky J: A note on genetic algorithms for large-scale feature selection. Handbook of Pattern Recognition and Computer Vision. Chen CH, Pau LF, Wang PSP (ed): World Scientific Publishing, Singapore; 1993. 88-107. 10.1142/9789814343138_0005
22. Moro S, Rita P, Cortez P: Bank Marketing [Dataset]. UCI Machine Learning Repository. (2014). https://doi.org/10.24432/C5K306.
23. Han J, Kamber M, Pei J: Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann, Waltham, MA;

2011. 10.1016/C2009-0-61819-5
24. He H, Garcia EA: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009, 21:1263-1284. 10.1109/tkde.2008.239
25. Cortes C, Vapnik V: Support-vector networks. Machine Learning. 1995, 20:273-297. 10.1007/bf00994018
26. Cover T, Hart P: Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967, 13:21-27. 10.1109/TIT.1967.1053964
27. Kohavi R: A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conferences on Artificial Intelligence. 1995, 14:1137-1145.
28. Maron ME: Automatic indexing: An experimental inquiry. Journal of the ACM. 1961, 8:404-417. 10.1145/321075.321084
29. Breiman L: Random forests. Machine Learning. 2001, 45:5-32. 10.1023/a:1010933404324
30. Powers DMW: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. 2011, 2:37-63. 10.9735/2229-3981
31. Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J: Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society. 2003, 54:627-635. 10.1057/palgrave.jors.2601545
32. Yeh IC, Lien CH: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications. 2009, 36:2473-2480. 10.1016/j.eswa.2007.12.020
33. Larose DT: Data Mining Methods and Models. John Wiley & Sons, Hoboken, NJ; 2006. 10.1002/0471756482
34. Weiss GM: Mining with rarity. ACM SIGKDD Explorations Newsletter. 2004, 6:7-19. 10.1145/1007730.1007734
35. Dietterich TG: Ensemble methods in machine learning. Multiple Classifier Systems. Springer, Berlin, Heidelberg; 2000. 1857:1-15. 10.1007/3-540-45014-9_1
36. Breiman L: Bagging predictors. Machine Learning. 1996, 24:123-140. 10.1007/bf00058655
37. Quinlan JR: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA; 1993.