https://doi.org/10.1007/s10796-024-10559-x

Features Selection for Credit Risk Prediction Problem

Ines Gasmi[1] · Sana Neji[2] · Salima Smiti[1] · Makram Soui[2]

Accepted: 10 November 2024

Abstract

Credit risk assessment has drawn great interests from both researcher studies and financial institutions. In fact, classifying an applicant as defaulter or non-defaulter customer helps banks to make a reasonable decision. The classification of applicants is based on a set of historical information of past loans. Data sets for analysis may include different features, many of which may be irrelevant to the decision making process. Keeping irrelevant features or leaving out relevant ones may be harmful, causing generation of poor quality patterns that may lead to confusion decision. Determining an appropriate set of predictors is an important challenge in credit risk prediction research which guarantees better decision-making. It is the task of searching the smallest subset of features that provide the highest accuracy and comprehensibility. Thus, this study proposes feature selection-based classification model on credit risk assessment. To this end, five algorithms are applied, Speed-constrained Multi-objective PSO (SMPSO), Non-dominated Sorting Algorithm (NSGA-II), Sequential Forward Selection (SFS), Sequential Forward Floating Selection (SFFS), and Random Subset Feature Selection (RSFS). The selected subset is evaluated based on three classifiers K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN). Our proposed model is validated using three real-world credit datasets. The obtained results confirm the efficiency of SMPSO-KNN model to select the most significant features and provide the highest classification accuracy compared to existing models.

Keywords Feature selection · Credit risk classification · Multi-objective optimization

1 Introduction

Evaluating credit risk is being paid considerable attention following the financial crisis in the world. In this way, various credit risk assessment models have been proposed for making better decision-making (Doumpos et al., 2019; Ferreira et al., 2019; Guo et al., 2016; Soui et al., 2019). Credit risk assessment is a risk prediction which aims to identify and classify applicants' status. The idea is to use a credit decision model which involves a better trade-off between risk recognition, minimization, and presentation. Better risk recognition is achieved with a preferment and accurate classification technique. Whereas a better risk minimization need the use of the most significant historical data in the decision process. Finally, a better risk presentation is reached by using an understandable model. It is the main factor that facilitates prediction and presents a robust decision-making system.

Different challenges make credit risk classification a difficult task. Classification aims to predict a target class based on a combination of some information or features. This problem has been tackled using various statistical and artificial intelligence techniques that are applied to automatically reveal knowledge from a given dataset. Datasets include of a set of features that can be considered as independent attributes that may be important for the decision and the learning process. However, building a classification model based on the full dataset can reduce accuracy because data often contain irrelevant and redundant data. Also, using a large number of features can cause a high-dimensional problem. It often leads

✉ Makram Soui
msoui@umich.edu

Ines Gasmi
gasmioines@gmail.com

Sana Neji
neji@umich.edu

Salima Smiti
salima.smiti92@gmail.com

[1] LARIA, National School of Computer Science, Manouba, Tunisia

[2] University of Michigan-Flint College of Innovation & Technology, Flint, Michigan, USA

Published online: 03 January 2025

to increase the complexity of the search space; as a result, the performance of the classifier decreases, the memory storage gains more, and the computational cost increases.

In this way, various dimensionality reduction methods have been proposed. According to Li et al. (2017) dimensionality reduction is divided into two main techniques: (1) Feature Selection (FS) and (2) Feature Extraction (FE). The feature selection is to determine a subset of relevant and non-redundant features from the original ones. While feature extraction remodels the original dataset by transforming the initial feature space into new features subspace with low-dimensionality using linear combinations. These combinations of features do not preserve the physical meaning of features which make them less interpretable, and the information regarding the importance of the original features are often hidden. Thus, FS has become a mandatory task in machine learning to eliminate the redundant and irrelevant features from the fast-increasing amount of data in the real world. The aim of feature selection is cleaning and filtering the dataset to improve the performance of data mining and machine learning techniques and enhance the comprehensibility of the proposed model.

In this way, many features selection techniques have been applied (Li et al., 2017; Mafarja et al., 2019; Wang et al., 2018; Mafarja & Mirjalili, 2018). These techniques are divided into three categories. A Filter-based model is applied independently before the learning process based on such ranking metrics. According to Zhang et al. (2015) and Su and Lin (2011), the effectiveness of this model relies mainly on the dataset itself rather than on the classifier. In literature, various filter algorithms have been used such as principal component analysis (Katrutsa & Strijov, 2017; Song et al., 2017), information gain (Alhaj et al., 2016; Koutanaei et al., 2015), relief (Urbanowicz et al., 2018), etc. A Wrapper based model which requires the performance of a learning algorithm to select a subset of features. This dependence on the classifier accuracy is considered as the vital advantage of this model and makes it preferred more than the filter method. Wrapper models are building also using different algorithms as genetic algorithm (Oreski & Oreski, 2014), genetic programming (Tran et al., 2016), practical swarm optimization (Xue et al., 2013), and artificial bee colony (Hancer et al., 2018), etc. And, finally, embedded based model that integrates the feature selection algorithm in the learning process, and considered it as an element in the classifier.

In the field of credit risk assessment, extensive FS methods have been proposed (Chen et al., 2016; Jain et al., 2018; Rao et al., 2019). However, there are few studies which propose feature selection approaches which can significantly improve the risk prediction rates. The main goal of FS techniques is to minimize irrelevant or redundant features, and select the significant ones and therefore maximizing the credit risk prediction and minimize the computational time. Most traditional search approaches have been used (e.g. Sequential Search, Sequential Floating and Search, filter approaches, etc.). However, these approaches fail to local maxima and minima, and hardly applied to solve real-world applications with multi-objectives, such as credit risk prediction (Huang et al., 2010). Thus, one of the possible approaches to deal with this issue is the use of Multi-Objective Optimization (MOO) algorithms. The existing studies (Ileberi et al., 2022; Kasongo, 2021; Rostami et al., 2021) define the FS step as a single-optimization problem using only one objective, i.e. accuracy rate. While multi-objective optimization algorithms aim to ameliorate the generalization efficiency in term of supervised classification and counterbalances the bias against the classes with limited samples in a dataset (Zhu et al., 2017). In addition, multi-objective FS algorithms determine simultaneously more than one feature subsets in the Pareto Front that have the same importance.

In traditional machine learning approaches, feature selection is often performed as a pre-processing step using filter-based methods or embedded techniques within the learning algorithm. However, these methods may not fully explore the complex relationships within the data or adequately address the high-dimensional nature of modern datasets. A clear benefit comes from search-based algorithms, such MOO approaches, which methodically search the feature space to identify ideal subsets that balance between reducing duplication and optimizing predictive performance. In contrast to traditional approaches, which could encounter difficulties with local optima and are not equipped to effectively manage multi-objective optimization, search-based algorithms offer a more comprehensive and adaptable structure for feature selection.

In this study, we adopted the SMPSO algorithm in order to minimize the number of features and maximize the accuracy of credit risk prediction. We compared our approach with another Multi-objective evolutionary algorithm NSGA-II and with three baseline algorithms: SFS, SFFS, and RSFS. NSGA-II was chosen due to its well-documented effectiveness in multi-objective optimization tasks, making it a suitable candidate for our study. Moreover, NSGA-II offers robustness and efficiency, aligning with our objective of maximizing feature selection performance in credit risk prediction. Its extensive use and established reputation in the literature further justified its selection as a representative genetic algorithm for our comparative analysis. In literature, SFS is one of the most effective FS algorithms in literature is SFS, which is distinguished by its ease of use and speed (Marcano-Cedeño et al., 2010). SFFS is also an effective wrapper variable selection algorithm. It is a floating variant of SFS considered as an intelligent search algorithm. The essential characteristic of SFFS algorithms is the use of starting rule when discovering the search space which minimizes the possibility of getting stuck in a local optimum. Finally, we applied the RSFS algorithm which tends to select the significant features by iteratively choosing random subsets of features from the full dataset that used by a set of classifiers. The effectiveness of the FS techniques is evaluated based on three popular algorithms: (1) Artificial Neural Network, (2) K-Nearest Neighbors, and (3) Support Vector Machine.

The rest of this paper is organized as follows. Section 2 presents the background of features selection and its importance in the credit risk prediction field. In Section 3, we define our proposed methodology. Then, an experimental study is presented in Section 4 to evaluate the performance of the proposed model. In Section 5, we discussed the related work. Finally, in Section 6, the conclusion and some perspectives are presented.

2 Background and Motivation

In financial institutions, credit risk is considered as a challenging task that aims to predict credit payment default. Inappropriate assessment of credit risk may lead a crisis in financial institutions. To this end, a large number of data mining and intelligent artificial techniques are applied in order to classify bank's applicants and distinguish between defaulter and non-defaulter customers. These techniques are mainly based on the analysis of bank applicants' history. A small improvement of 1% of the predictive accuracy of the detection of non-serious applicants will provoke an enormous gain for banks (Rao et al., 2019; Huang et al., 2010). Hence, data treatment is a key engine to build an accurate and interpretable classification model. In this way, feature selection methods are mainly used in order to eliminate redundant, irrelevant, and misleading features. This process improves simultaneously the predictive accuracy of the model and reduces the computational cost. Feature selection techniques can be used to achieve the best trade-off between:

1) **Accuracy:** FS aims to improve classification accuracy by using only the relevant and the most significant features for decision-making.
2) **Complexity:** Feature selection reduces the number of required variables which simplify and speed up the learning process.
3) **Interpretability:** FS eliminates irrelevant and redundant features which makes data more interpretable and, hence, improves the comprehensibility of the generated model.

Our research initially focused on the accuracy, complexity and interpretability, as goals. However, further analysis reveals potential conflicts among these objectives. Specifically, while reducing the number of features typically simplifies the model and enhances interpretability, it may not always lead to an improvement in accuracy. This discrepancy arises from the risk of losing information during feature selection, which could negatively impact performance. For example, in some cases when we try to improve the accuracy of the proposed model, this can hinder its complexity. Therefore our approach aims to explore these conflicts by providing real-world examples and scenarios in the field of credit risk assessment to illustrate the delicate balance between accuracy, complexity and interpretability. To this ends, we used Pareto optimality to address these conflicting goals and enhance our feature selection process.

Reducing the number of features can simplify the model and potentially improve accuracy by removing noise and irrelevant information. However, it may also lead to oversimplification. Complex relationships between features may exist, and by removing them, the model may fail to capture important patterns in the data, which would ultimately lower its predictive accuracy. By oversimplifying the model and selecting only a subset of features, these critical relationships may be overlooked, resulting in decreased predictive accuracy. For instance, omitting variables such as employment history or loan purpose may overlook important predictors of creditworthiness, ultimately reducing the model's accuracy.

Also, in credit risk assessment, correlation between demographic variables, such as age and income, and credit history variables, such as payment history and outstanding debt, may contribute to the complexity of the model. Balancing complexity with interpretability means selecting features that maintain predictive power while remaining understandable to applicants. Achieving this balance ensures accurate risk assessment while maintaining model transparency.

In the realm of multi-objective optimization for credit risk assessment, three conflicting objectives emerge: accuracy, complexity, and interpretability. Accuracy necessitates maximizing predictive performance by including all relevant features, thereby reducing the risk of oversimplification and ensuring robust risk assessment. However, complexity arises from the complicated correlation between demographic variables (e.g., age, income) and credit history indicators (e.g., payment history, outstanding debt), which can lead to model complexity and computational inefficiency. Simultaneously, interpretability requires simplifying the model to make it understandable to applicants, potentially compromising predictive accuracy by excluding essential predictors. Thus, the challenge lies in finding a delicate balance between accuracy, complexity, and interpretability to create a feature selection framework that effectively manages these conflicting objectives and generates reliable credit risk assessment models.

Generally, the full datasets used in credit risk assessment have high dimensional features. Indeed, irrelevant features in a training dataset could produce less accurate results in the classification analysis. In this context, feature selection is required to select the most significant features to increase the predictive accuracy, speed, and scalability. Thus, feature selection is considered as an important pre-processing step for building accurate classification models (Mafarja et al., 2019; Zhu et al., 2017; Marcano-Cedeño et al., 2010). It is used to identify and eliminate useless, trivial, and redundant features that do not contribute to the certainty of a classification model. On the another hand, feature selection is the process which allows reducing the feature space by selecting a subset of features (Yu & Liu, 2004). The Feature selection process takes the full training dataset as input and generates a subset of interesting features as output. This process includes three main steps: (1) subset generation, (2) subset evaluation, and (3) stopping criteria.

As shown in Fig. 1, feature selection methods are typically based on two requirements: search strategy and objective function (Zhang et al., 2014). The search strategy aims to select the candidate subset of features for searching all possible subset sequences. While, the objective function evaluates the derived subsets in terms of their goodness value. This value will be used then by the search strategy to choose the next candidate's solutions. The objective function is divided into three main categories: (1) filter methods, (2) wrapper methods, and (3) embedded methods.

Filter methods are mainly based on a certain statistical measure, such as correlation, F-score, Information Gain, Principal Component Analysis, etc., which are used to rank features in terms of their importance value. These methods are applied before the classification process to filter out the irrelevant features without considering any learning algorithm. Filter methods are classified into two categories: (1) multivariate methods and (2) univariate methods. Univariate methods aim to evaluate each feature independently, while multivariate methods are characterized by their ability to find a correlation between the features. Several Filter methods have been used such- as Pearson correlation coefficient (Grabczewski & Jankowski, 2006; Guyon & Elisseeff, 2003), mutual information (Hoque et al., 2014; El Akadi et al., 2008), chi-square test scores (Jin et al., 2006), etc.

The wrapper methods require a machine learning algorithm to evaluate each derived features subset during the selection process. The subset which provides the highest classification accuracy will be selected. Then, the same learning algorithm is used to build a predictive model. These methods are mainly based on search strategy process such as sequential methods, Genetic Algorithm, Genetic Programming, Branch and Bound method, Artificial Bee Colony, etc. These algorithms try to generate the relevant subset. However, they are very computationally cost, especially, for the large dataset. The selection of an effective search strategy technique is one of the crucial keys to obtain an optimal subset of features. Thus, sequential methods are one of the widely used wrapper methods. They start with an individual feature subset and alternatively add or eliminate features until reaching the stop criterion. The sequential methods are classified into two categories: Sequential Forward Selection (SFS) which begins with an empty set and tries to add relevant features. While Sequential Backward Selection (SBS) starts with the full set and removes the irrelevant features.

Finally, the embedded methods integrate the FS step as a part of the classifier. These methods often use decision tree algorithms, such as, C4.5 and CART. Embedded methods consist of determining only the weights of features or ranking them rather than providing a subset of features. These methods spend less executive time than the sequential method; however, they require direct adjustment of the learning process.

In the field of credit risk assessment, the prediction accuracy is strongly affected by the set of used features and the model's input data. Therefore, selecting the optimal feature subset for the decision process is vital for banks. According to [37,38], in financial risk prediction two essentials steps must be taken in consideration:

1. Finding the relevant and appropriate subset of features based on a robust FS technique.
2. Applying a classification algorithm in order to generate a credit risk assessment model using the selected features.

Fig. 1 Features selection requirements

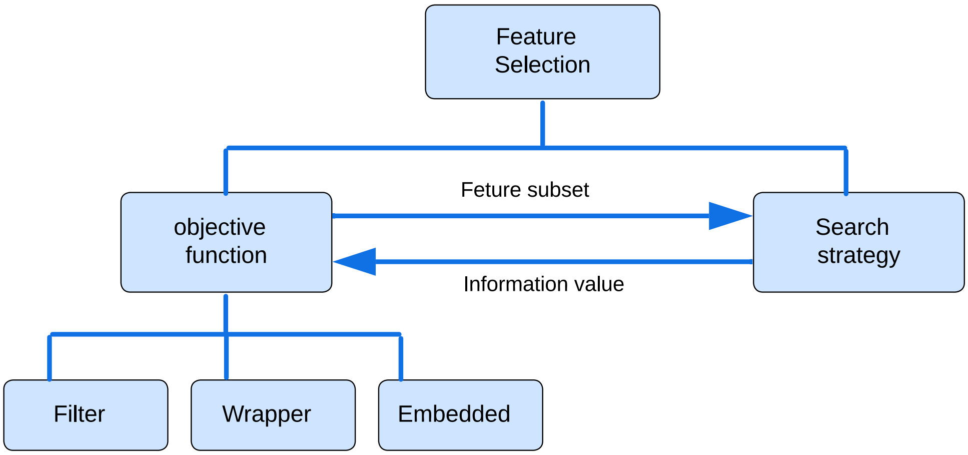
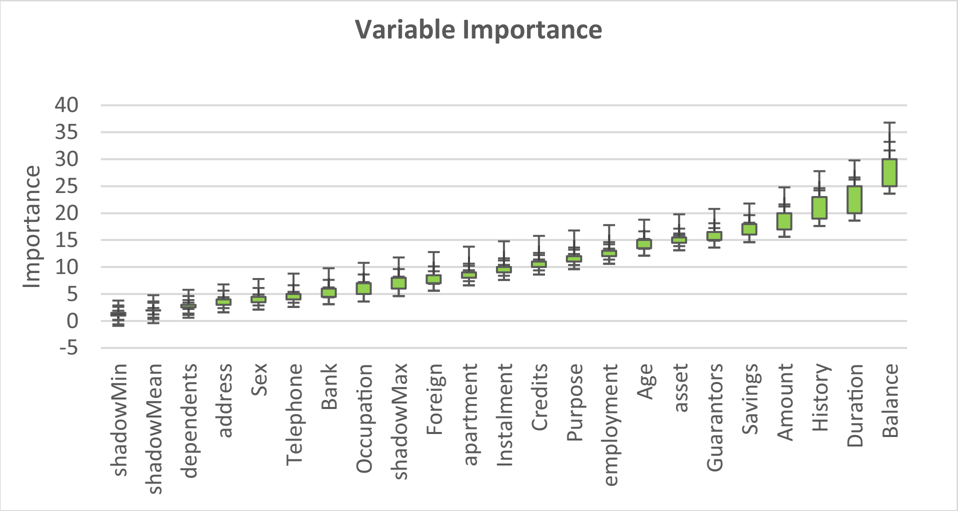


Fig. 2 Feature importance variation for German dataset



Classifying bank applicants based only on the most significant variables improves, simultaneously, the correctness of the generated model and its interpretability. However, there is no consensus about the features needed for the decision process. For instance, we utilized variable importance testing to assess the significance of features in the dataset. In fact, we aimed to demonstrate the varying importance of features, depending on the specific studied problem. Subsequently, we generated a boxplot diagram, as depicted in Fig. 2, to visually represent the distribution of feature importance scores. Notably, our analysis revealed that only 11 variables were considered significant for decision-making, including other current credit, purpose, present employment since, age, available assets, guarantors, savings stocks, credit amount, previous credit history, credit duration, and account balance. These variables were identified as crucial for accurate credit risk assessment. Conversely, the remaining features (telephone, current address, marital status, occupation, credit at this bank, foreign worker, instalment rate, housing) were considered non-significant for the decision-making process. The utilization of variable importance testing and the subsequent visualization through the boxplot diagram provided valuable insights into the relative importance of features, further reinforcing the necessity for a robust feature selection model to ensure the effectiveness and accuracy of the credit risk prediction model. This uncertainty about feature importance requires a robust feature selection model to ensure:

1. The minimization of overfitting: minimize redundant and missing data value means reducing the possibility of making decisions based on noisy data.
2. The maximization of accuracy: The aim is to build an accurate predictive model able to accurately classify applicants.
3. The minimization of training time: minimizing data variable allow reducing classifiers complexity and therefore speed up the training process.

3 Methodology

The aim of this study is to use a Multi-objective optimization algorithm SMPSO to select the most relevant subset of features that ensure the best trade-off between the previously mentioned objectives. To this end, ANN, KNN and SVM are used to evaluate the solutions generated by the SMPSO algorithm during the successive generations and to classify applicants. To evaluate the proposed model, we conducted a comparative study using NSGA-II and three baseline wrapper feature selection algorithms SFS, SFFS, and RSFS (As shown in Fig. 3). Wrapper method is a feature selection technique where the selection process is embedded within the model evaluation loop. In each iteration, the FS algorithm selects a given possible subset which includes the important features. Then, the classification algorithm uses this subset to build a proposed model. Next, the efficiency of this model is evaluated based on cross-validation method. Finally, the optimal subset that yields the best accuracy is selected. In this study, our individual is represented as a simple coding scheme where we use a binary chromosomes representation. Each attribute is represented as zero or one. These individual are evaluated based on two objective functions which determine the importance of the obtained subset of features using the following equation 1:

$$f(x) = \begin{cases} \text{Minimize } f_1(x) = NF(S_i) \\ \text{Maximize } f_2(x) = W(S_i) \end{cases} \tag{1}$$

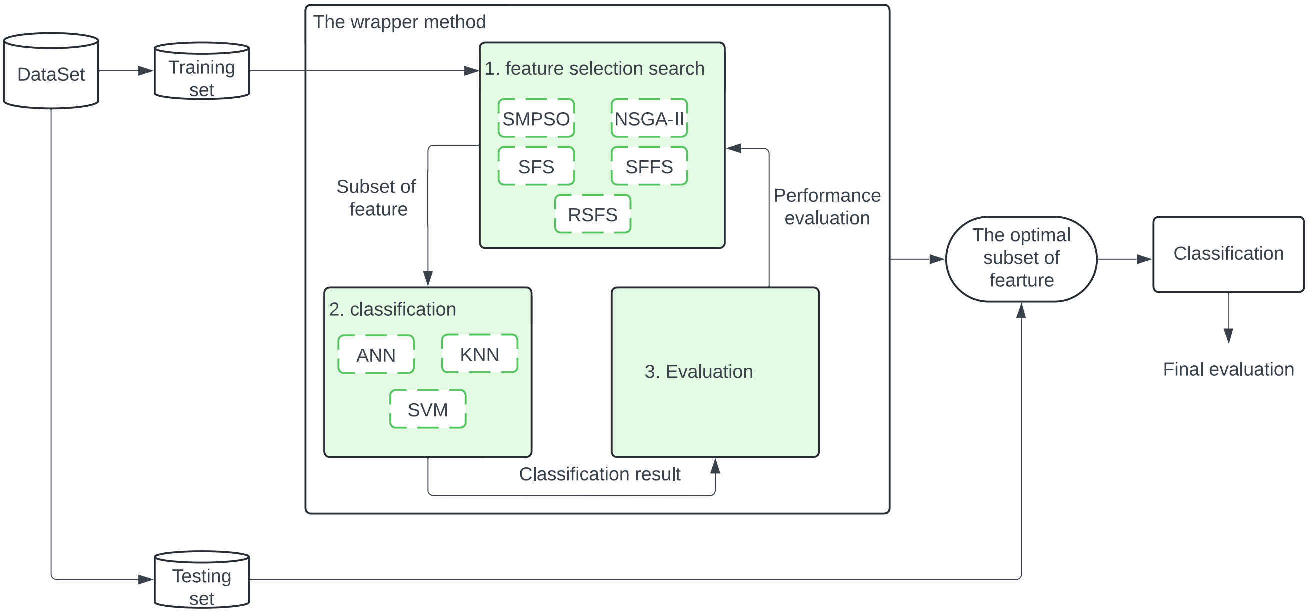


Fig. 3 The proposed wrapper feature selection model

The first objective $f_1(x)$ consists of minimizing the number of features in the generated subset S_i. The optimization process aims to reduce the complexity of the input variables by minimizing the number of features for the classifier.

The second objective $f_2(x)$ tends to maximize the weight of the selected features. To this end, the Information Value (IV) is used to select the relevant features. IV ranks features in term of their importance. It is widely applied to determine the importance of features in the credit scoring application such as the probability of default prediction. The IV aims to select features having an impact on the credit risk decision. Thus, a weight w is affected to each feature. The weight values are ranged between -1 and 1. The IV is measured to distinguish between features having highly or slightly relationship with class label. The features that are highly correlated with the class label have an impact on credit risk, while, the remaining features have a slight impact on risk decision. IV measures the difference between the percentage of 'goods' and the percentage of 'bad' multiplied by the WOE for each respective attribute. The IV is measured as follows:

$$IV_i = \sum_{i=1}^{n} (G\% - B\%) \times WOE_i \tag{2}$$

Where:

$$WOE = \ln\left(\frac{g_i/G}{b_i/B}\right) \tag{3}$$

IV_i: the weight of each feature in the obtained subset and n is the size of the subset.

G: the total number of good instances.
B: the total number of bad instances.
g_i: the number of good within the feature i.
b_i: the number of bad within the feature i.

3.1 Features Selection Algorithms

3.1.1 SMPSO

Speed-constrained Multi-objective PSO is an improved variant of MOPSO algorithm (Coello & Lechuga, 2002) proposed by Nebro et al. (2009). As repsented in Fig. 4, SMPSO algorithm aims to calculate the velocity of particles based on constriction mechanism. It uses an external archive to save the approximate non-dominated solutions during the search process. This variant integrates a particular leader's archive for storing leaders (particles which are considered as personal best). The upper size of this archive is limited by the number of particles in the swarm. If the archive size limit is reached the crowding distance of NSGA-II is used to determine which particles should be removed to promote diversity and to maintain the fixed size. SMPSO is a population-based algorithm, like the traditional PSO, it performs by manipulating a set of individuals "particle" which known collectively as the "swarm". In order to imitate the swarm behavior and explore the search space, each particle is permitted to move (fly) towards the optimal solution. SMPSO is suggested to optimize several objectives function. The position of particle xi at a given generation t is upgraded based on the following equation:

$$\mathbf{x}_i(t) = \mathbf{x}_i(t-1) + \mathbf{v}_i(t) \tag{4}$$

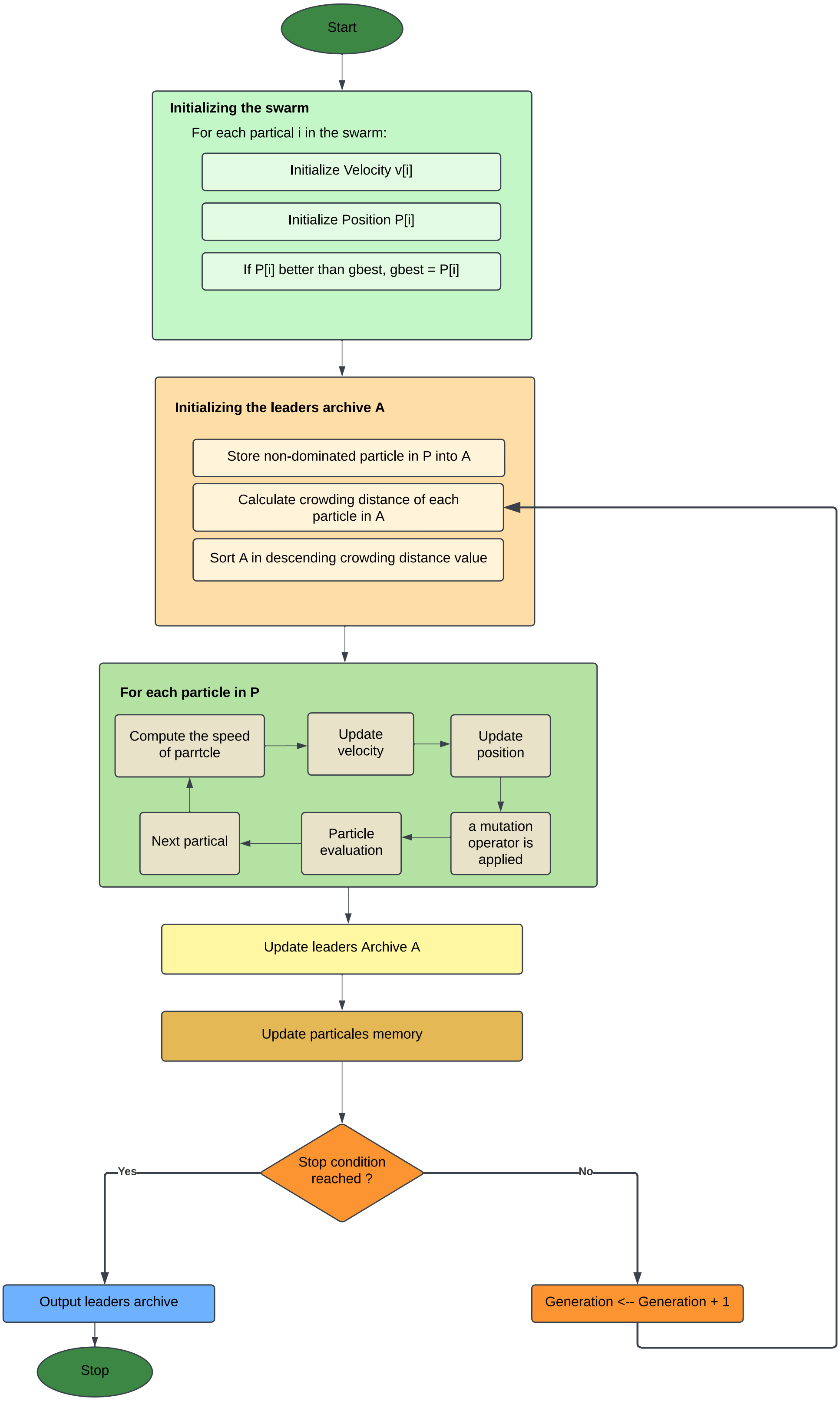

Start
Initializing the swarm
For each partical i in the swarm:
Initialize Velocity v[i]
Initialize Position P[i]
If P[i] better than gbest, gbest = P[i]
Initializing the leaders archive A
Store non-dominated particle in P into A
Calculate crowding distance of each particle in A
Sort A in descending crowding distance value
For each particle in P
Compute the speed of parrtcle
Update velocity
Update position
a mutation operator is applied
Particle evaluation
Next partical
Update leaders Archive A
Update particales memory
Stop condition reached ?
Yes
No
Output leaders archive
Generation <-- Generation + 1
Stop

Fig. 4 Flowchart of SMPSO algorithm

Where $v_i(t)$ denote the velocity of particle I that speed the optimization process and influence both the cognitive (personal) experience knowledge and the social (global) experience knowledge from the all particles, it is defined as follows:

$$\mathbf{v}_i(t) = w \cdot \mathbf{v}_i(t-1) + C_1 \cdot r_1 \cdot \left(\mathbf{x}_{p_i} - \mathbf{x}_i\right) + C_2 \cdot r_2 \cdot \left(\mathbf{x}_{g_i} - \mathbf{x}_i\right) \quad (5)$$

Where:
i: denote the particle index.
w: is the inertia coefficient.
$v_i(t-1)$: particle velocity in the previous generation.
x_pt: is the local best.
x_gt: is the global best or leader.
$C1, C2$: represent the acceleration coefficient which controls respectively the cognition component (personal) and social component global best particles.
$r1, r2$: are random values in the range of [0.1] which varies in each velocity update.

To determine particles' velocity, SMPSO applies a constriction coefficient inspired from the constriction factor χ originally proposed by Clerc and Kennedy (2002). The coefficient is defined as follow:

$$\chi = \frac{2}{2 - \varphi - \sqrt{\varphi^2 - 4\varphi}} \quad (6)$$

Where:

$$\varphi = \begin{cases} C_1 + C_2 & \text{if } C_1 + C_2 > 4, \\ 0 & \text{if } C_1 + C_2 \le 4. \end{cases} \quad (7)$$

In addition, SMPSO bounds the accumulated velocity of each variable j (in each particle) using of the following velocity constriction equation:

$$v_{i,j}(t) = \begin{cases} \text{delta}_j & \text{if } v_{i,j}(t) > \text{delta}_j \\ -\text{delta}_j & \text{if } v_{i,j}(t) \le -\text{delta}_j \\ v_{i,j}(t) & \text{otherwise} \end{cases} \quad (8)$$

Where:

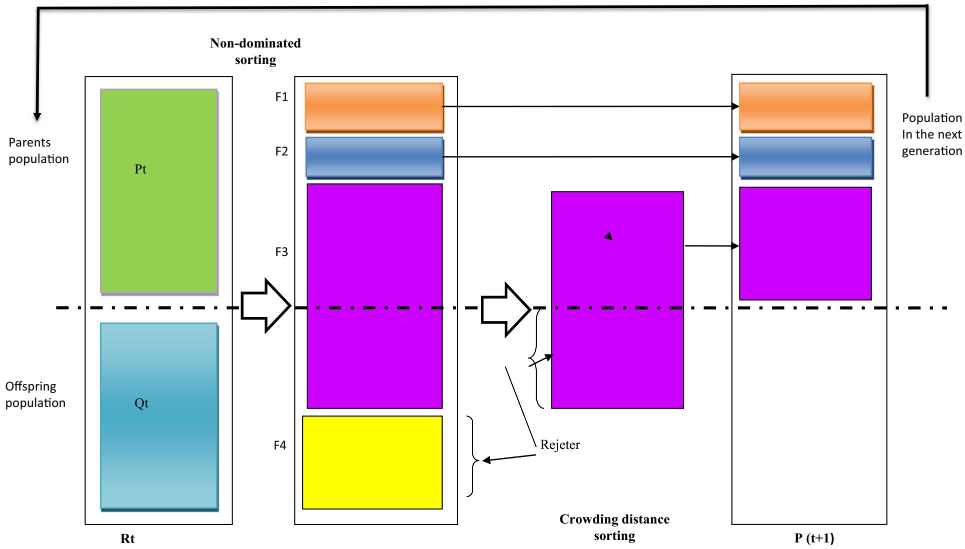
$$delta_j = \frac{\left(upper_limit_j - lower\_limit_j\right)}{2} \quad (9)$$

3.1.2 NSGA-II

NSGA-II is a Pareto Front based algorithm which aims to search the set of non-dominated solutions in a single run based on three main principles: (1) fast non-dominated sorting procedure which ranks solutions into different front based on the level of non-domination. (2) Crowding distance operator which calculates the distance between each solution and its neighbors. This operator tends to preserve better diversity by maximizing the crowding distance average. (3) Finally, the elitism principle which enhances the convergence characteristics of a multi-objective evolutionary algorithm because it stores all non-dominated solutions. Elitism aims to speed up the performance of the algorithm and to save the best solutions once they are found. In this work we will study the efficiency of NSGA-II to generate the best set of relevant features. As represented in Fig. 5, NSGA-II includes the following steps:

1. The generation of the initial population. Indeed, the population P_0 is initialized randomly with a fixed population size.
2. The evaluation of the first population based on the fitness functions.

Fig. 5 NSGA-II overview



3. This step sorts all the solutions of P_0 into different non-dominated layer using the Fast non-dominated-sort procedure.
4. Next, crossover and mutation operators are applied and the offspring population is selected.
5. Once the children population is produced, all the obtained individuals are evaluated based on the objective functions.
6. Then, the offspring and P_0 population are combined into a new population Pt.
7. All the solutions of this population (Pt) are ranked in order to obtain the different fronts.
8. Then, the next population is generated. To this end, the half top-ranked function is used to select best solution. When the half is belonging inside a front F_i. Solutions of this rank should be stored using the crowding distance, and those which have greater crowding distance values will be selected.
9. Finally, the crossover and mutation procedures are applied to produce the new populations.
10. The algorithm finished the work if the stopping condition is Therefore our approach aims to explore these conflicts reached; otherwise, it turned back to step 2.

3.1.3 Sequential Forward Selection (SFS)

Sequential Forward Selection is a well-known feature selection algorithm (Pohjalainen et al., 2015). It started with an individual feature which has the highest significance value, and progressively adds the important feature from the original dataset. These additional features are evaluated based on a criterion function that should be optimized when including a new feature using a classification algorithm. SFS updates the selected features by recurrently based their importance. In each iteration, one feature will be added to the subset the selected of features when the classification accuracy is increased. SFS generates a new robust subset which is composed of the most efficient features and generates less classification error.

3.1.4 Sequential Forward Floating Selection

Sequential Forward Floating Selection is considered as a variant of SFS algorithm proposed by Reunanen (2003). In SFS algorithm there is no backward step, when a feature is added to the final features set, it cannot be removed. Thus, one of the main issues of the SFS algorithm is its monotonic growing feature set. To this end, SFFS algorithm is proposed in order to allow removing already selected features. First, SFFS uses the SFS algorithm in order to generate an initial feature set. Similar to SFS, the set of features is created by adding features based on a criterion function. In the next iterations, each feature which is able to improve this criterion is excluded. However, in some iteration, the feature set can be reduced by eliminating the least significant one. In this floating variant, the algorithm continues to increase or decrease the number of features until the optimal subset of feature is determined.

3.1.5 Random Subset Feature Selection

Random Subset Feature Selection is a feature selection algorithm that aims to find the optimal subset of features which can contribute to boosting the efficiency of the classifier. These features are determined by repetitively fed a classifier with a random subset of features as input. The importance of each feature is adjusted based on the classification performance of the chosen subset. In greedy methods such as SFS and SFFS, the importance of a feature is directly calculated by including or excluding it from the created subset. However, RSFS evaluates each feature regarding its average usefulness with other feature combinations. In other words, the relevance of every single feature is assessed regarding its contribution in correct classification. RSFS iteratively classifies a random subset of features many times as it is necessary to determine the relevant features which commonly seem useful according to the random components of the process.

3.2 Machine learning algorithms

3.2.1 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (KNN) is one of the simplest supervised machine learning algorithms which is often used to solve classification problems (Kramer, 2013). It classifies a candidate data point based on how its neighbors are classified based on Euclidian distance. At the end of the classification process, KNN assigns the class of the current point based on the majority of the vote between K nearest neighbor's classes, where K denotes the number of nearest neighbors. In order to achieve a high accuracy, it is important to thoroughly choose the optimal value of K. For example, if $K = 1$, then the data point is simply assigned to the class of its single nearest neighbor. Several studies confirm that there is no optimal value of K that can be suitable for all the datasets since each one has its specific requirements (Pohjalainen et al., 2015; Imandoust & Bolandraftar, 2013; Wang et al., 2007). Commonly, data scientists suggest that the K value could be an odd number when the number of classes is 2.

3.2.2 Support Vector Machine

Support vector machine is one of the most popular supervised learning algorithms that support classification and regression (Cortes & Vapnik, 1995). However, it is well used to

Table 1 Characteristics of three real datasets

Name	Instance	Pred. attr.	Good cred.	Bad cred.	No. of classes
German	1000	20	700	300	2
Australian	690	12	307	383	2
Taiwan	30000	24	23364	6636	2

address classification problems. SVM aims to find the best separating hyperplane between two classes based on mapping the pattern points into high dimensional space that is performed based on a set of mathematical functions called Kernels. SVM generates a linear decision boundary based on the largest distance which called maximum margin. In other words, the basic idea of SVM is to find the optimal separating hyperplane which maximizes the margin of separation between support vectors of the two determined classes.

3.2.3 Artificial Neural Network

An Artificial Neural Network is a computational model inspired by the structure of the human brain that can be used as a robust technique in various domains (Kruppa et al., 2013; Kang et al., 2010). ANN is capable of learning from patterns and trying to figure out a function or a set of calculations that ultimately generates useful results related to a given classification problem. ANN is a highly structured model based on a set of layers. The first layer defines the input layer where the neural network takes a set of features as inputs. While, hidden layers consist of performing complex calculations. Then, the output of a given layer will be forward to the next hidden layer. The last one is the output layer which provides the predicted class labels.

4 Experimental Setup

In order to evaluate the efficiency of the proposed model, this section addresses four main research questions and explains how our experiments are designed to address them. The objective of this study is to investigate the benefits of the feature selection technique to improve the effectiveness of classifiers for predicting credit risk.

4.1 Dataset

To assess the advantage of features selection algorithms, this study used three benchmark datasets: (1) German[1] , (2) Australian [2] and (3) Taiwan[3] dataset . These datasets are obtained from the University of California, Irvine (UCI) Machine Learning Repository. German dataset contains 1000 instances in which 700 samples are good applicants, and 300 are bad applicants. For each applicant, 20 predictive attributes are used. The Australian dataset includes 690 observations with 14 attributes, where 303 instances present good applicants and 387 present bad applicants. Finally, Taiwan dataset is composed of 30000 instances, where 23364 are good applicants and 6636 are not. Each observation has 24 predictive attributes. A summary of the main characteristics of these three datasets has been presented in Table 1. The used datasets are randomly separated into two parts (training and testing partitions). In this work, we use 70% of the dataset for training the classifiers whiles the remaining 30% for testing the proposed model.

4.2 Parameter Setting

First, we need to set out the algorithms parameter because it significantly influences the performance of search algorithms. In order to guarantee an effective convergence of SMPSO and NSGA-II, it is important to carefully consider their parameter setup. Thus, we calculate and compare the p-value of the obtained results given by SMPSO and NSGA-II algorithms to statistically determine the results significance difference. For SVM, the cost of classification C and the kernel parameter (gamma) take different values as input and returns the best value for classification. For KNN, the number of neighbors (K) is the most important parameter to improve classification accuracy. The optimal value of K provides the lowest test error rate. For the NN algorithm, we adjusted various parameters to optimize the model: the number of epochs, the number of hidden layers, and the number of neurons in each layer, the learning rate and finally the momentum rate. The Table 2 shows the important setting parameters for the used algorithms. In this study, we used the 10-fold cross-validation to evaluate the obtained results.

4.3 Research Questions

We designed our experiment to answer the following four research questions:

RQ1: To what extent can the studied feature selection algorithms (SMPSO) improve the efficiency of the proposed classifiers?

RQ2: To what extent the feature selection algorithms able to reduce the execution times of the studied classifiers?

[1] https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

[2] https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval

[3] https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Table 2 Parameter settings for the used algorithms

Algorithms	Parameters
SMPSO	Swarm Size: 100
	Mutation: polynomial mutation, pm = 1.0/L
	Archive leaders size: 100
NSGA-II	Population size: 100
	Selection: Binary tournament selection
	Crossover: Single point crossover, pc = 0.9
	Mutation: polynomial mutation, pm = 0.1
SVM	Kernel function: RBF kernel
	C(cost): [0.1, 0.5, 1, 2, 5, 10, 20, 50];
	Gamma: [0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10];
KNN	k (NumNeighbors): 1, 2, 3, ... N
	K-fold: 5
	Mean squared: true
	Distance: 1
ANN	Number of epochs: 20
	Number of hidden layers: 1
	Number of hidden units: 100
	Momentum rate: 0.3
	Learning rate: 0.5

RQ3: To what extent the selected features and the classification results are statistically significant?
RQ4: How does the proposed model performs compared to existing ones that use feature selection algorithms?

To answer RQ1, we evaluate the efficiency of the used classifiers: SVM, ANN, and KNN in terms of accuracy and F1-score measures based on the subset of features generated by five FS algorithms: SMPSO, NSGA-II, SFS, SFFS, and RSFS. The accuracy metric corresponds to the ratio of the correctly classified applicants (good and bad) on a particular dataset defined in equation 10. The F1-score is used also to evaluate the obtained results. As shown in equation 11, F1-score determines the average between the sensitivity and the specificity, where the sensitivity indicates the percentage of correctly classified good loans and the specificity represents the percentage of correctly classified bad loans.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{10}$$

$$F_1 - \text{score} = 2 * \frac{\text{sensitivity} * \text{specificity}}{\text{sensitivity} + \text{specificity}} \tag{11}$$

Where:

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{13}$$

To answer RQ2, we study the impact of the feature selection algorithms on the execution time of the studied classifiers. The execution time or the computational time (ICT) is an important measure used to evaluate the performance of classifiers. To this end, we compared the execution time of the classifiers before and after applying the feature selection techniques.

To answer RQ3, we statistically examine the performance and significance of our proposal based on two statistical tests: (1) paired t-test and (2) Pearson chi-squared test. First, t-test is used to compare the accuracy of our proposed against the other studied features selection algorithms. In this way, we test the following hypotheses:
H0: the three used learning algorithms have equal classification accuracy with the used FS algorithms.

H1: The SMPSO-KNN algorithm provides an improved result compared to the other applied classifiers and FS algorithms.

Second, the chi-squared test is used to determine the relationship between the selected features (obtained by SMPSO algorithm) and the credit risk. In this case, the p-value threshold is considered as a critical metric for the dependence or the independence of each feature and credit risk.

To answer RQ4, we compare the results obtained by SMPSO-KNN model with other similar works that use the same datasets (German, Australian, and Taiwan datasets). This comparison is based on two evaluation criteria: accuracy and F1-score.

4.3.1 Results for Research Question 1

In order to build an accurate credit risk assessment model, a feature selection step is applied. It is an important step in data preprocessing that chooses the best features in order to improve the classification accuracy and decrease running time. In this experiment, we compared the proposed SMPSO based model with four feature selection algorithms, NSGA-II, SFS, SFFS, and RSFS. The proposed models are implemented and developed using Matlab 2017b platform. Table 3 represents the sub-set of the most significant features selected by SMPSO, NSGA-II, SFS, SFFS, and RSFS. We observe that SMPSO has the least number of features when compared to the other discussed algorithms with all the tested datasets. For the German and Taiwan datasets, the number of features is reduced to only 6 important ones. While for the Australian dataset, we notice that the SMPSO selects 5 attribute from the input variables.

In this study, the effectiveness of these obtained subsets is evaluated based on ANN, KNN, and SVM classifiers. To reveal the impact of feature selection algorithms on the classification accuracy, we run these classifiers with and without the FS process, the obtained results is presented in Tables 4

Table 3 Features selected by various feature selection algorithms

Dataset	Feature Selector	Selected Features	No of features
German dataset	SMPSO	7,3,10,6,5,13	6
	NSGA-II	13, 1, 3, 5, 7, 9, 11	7
	SFS	3, 4, 6, 7, 19, 13, 8	7
	SFFS	19, 7, 8, 9, 15, 12, 14, 3	8
	RSFS	19, 13, 11, 15, 16, 3, 12, 9, 6, 14,18, 4, 8, 7	14
Australian dataset	SMPSO	2, 5, 4, 8, 3	5
	NSGA-II	10, 2, 5, 7, 4, 11, 6	7
	SFS-ANN	1, 3, 10, 13, 7, 4, 8	7
	SFFS	3, 10, 13, 4, 1, 8, 5, 11	8
	RSFS	3, 5, 11, 4, 1, 10, 7, 13, 8	9
Taiwan dataset	SMPSO	5, 10, 3, 12, 16, 22	6
	NSGA-II	3, 14, 5, 2, 7, 22, 9	7
	SFS	2, 5, 9, 14, 10, 17,12, 22	8
	SFFS	9, 14, 5, 10, 17, 7, 24	7
	RSFS	6, 23, 7, 12, 10, 5, 15, 14, 13, 18,2, 17, 24, 8, 16	15

and 5. We deduce that the accuracy and F1-Scores of ANN and KNN are ameliorated compared to the results achieved without applied feature selection algorithms. As shown in Table 4, we note that the SMPSO-KNN outperforms the other used algorithms. It provides the highest accuracy with an average of 95% for German dataset instead of 73% using the full dataset, 97% for Australian dataset instead of 75%, and 93% for Taiwan datasets instead of 76%. As depicted in Table 5, the proposed SMPSO-KNN model has the best F1-score rate for Australian and Taiwan dataset with 96%. However, for the German dataset, both SMPSO-KNN and NSGA-II-ANN models perform the best F1-score rate with

Table 4 The comparison of accuracy using feature selection algorithms

Method		German dataset	Australian dataset	Taiwan dataset
None	ANN	74	85	76
	KNN	73	75	76
	SVM	71	72	78
SMPSO	ANN	92	93	89
	KNN	95	97	93
	SVM	65	70	79
NSGA-II	ANN	91	94	90
	KNN	91	96	91
	SVM	72	70	72
SFS	ANN	83	78	83
	KNN	73	75	76
	SVM	79	76	80
SFFS	ANN	82	88	81
	KNN	72	75	79
	SVM	70	69	78
RSFS	ANN	75	86	78
	KNN	70	74	76
	SVM	71	70	79

Table 5 The comparison of F1-score using feature selection algorithms

Method		German dataset	Australian dataset	Taiwan dataset
None	ANN	77	86	57
	KNN	71	86	57
	SVM	74	82	53
SMPSO	ANN	92	94	95
	KNN	94	96	96
	SVM	69	82	52
NSGA-II	ANN	94	92	93
	KNN	92	93	89
	SVM	71	80	56
SFS	ANN	85	88	67
	KNN	78	85	60
	SVM	72	81	50
SFFS	ANN	83	89	58
	KNN	75	87	58
	SVM	73	80	52
RSFS	ANN	78	77	56
	KNN	70	80	57
	SVM	77	79	55

94%, when without FS KNN and ANN provide respectively 71% and 77% of F1-score. However, the performance of SVM degrades from 71% of accuracy to 65% when the FS is applied. The SVM is especially suitable for classification problems with a huge number of features (Jin et al., 2006). As FS ultimately contributes to minimizing the number of features, thus the use of SVM would be contradictory. These results confirm the effectiveness of this model and its ability to correctly classify and evaluate new applicants.

Although NSGA-II is certainly a very effective and adaptable optimization method, there are a number of important aspects that contribute to SMPSO's effectiveness in this particular field. Firstly, the feature selection process in credit risk assessment is a challenging optimization task where accuracy, interpretability, and computational efficiency must all be balanced. These requirements are well-suited to SMPSO's adaptability and speed-constrained mechanism, which allows it to efficiently explore the solution space while maintaining diversity among feature subsets. This adaptability is crucial in navigating the high-dimensional and complex datasets that are part of credit risk assessment, where standard optimization approaches may struggle to converge on optimal solutions.

Moreover, the nature of credit risk assessment datasets presents unique challenges that may not be fully addressed by conventional optimization algorithms like NSGA-II. These challenges include highly correlated features, noisy data, and imbalanced class distributions, all of which can impact the effectiveness of feature selection techniques. SMPSO can effectively solve these difficulties and find feature subsets that balance model complexity and predictive performance because of its dynamic adjustment of exploration and exploitation strategies based on the problem characteristics.

Additionally, the focus on maximizing the importance of selected features while minimizing their number aligns closely with the objectives of credit risk assessment, where it is critical to identify the most relevant risk factors Through the integration of these particular objectives into its optimization process, SMPSO can tailor its search to target feature subsets that are not just predictive but also comprehensible and useful for making decisions.

In summary, while NSGA-II is still a powerful optimization algorithm in many other fields, the particular requirements of feature selection in credit risk assessment required a method that improve its dynamic adjustment of search space exploration and find features subset that balance the model complexity and its predictive accuracy by giving equal weight to interpretability and accuracy. Therefore, SMPSO is a good choice for optimizing the feature selection process in a significant search area for credit risk prediction.

SMPSO employs a speed-constrained mechanism that allows particles to explore the solution space efficiently while maintaining diversity. As a result, SMPSO is able to discover a diverse set of feature subsets that represent different trade-offs between accuracy and complexity. Maintaining diversity is essential in feature selection, as it helps prevent premature convergence to suboptimal solutions and promotes the identification of novel, high-quality feature subsets. Additionally, the adaptive search mechanism of SMPSO allows it to adaptively adjust its exploration and exploitation strategies based on the characteristics of the optimization problem. In the context of credit risk assessment, where datasets may exhibit complex patterns and relationships, this adaptive search mechanism enables it to effectively explore the solution space and identify relevant feature subsets. Credit risk assessment datasets often contain a large number of features, which can present challenges for optimization algorithms. SMPSO is well-suited for processing high-dimensional data because of its effective exploration mechanism and speed-constrained methodology, which enable it to find significant features even in complex datasets.

4.3.2 Result of Research Question 2

To better evaluate our proposed model, it is important to study the impact of the feature selection algorithm in term of the execution time of classifiers. To this end, we executed the proposed models on a standards desktop computer (i5 CPU running at 2,6 GHz with 6 Go of RAM). The dimensionality reduction of the dataset can reduce the running time and speed up the algorithm execution. Figure 6 presents the execution time of the studied algorithms used in our experiments. The result shows that the execution time decreases when using the feature selection techniques. It is clear from this figure, that the SMPSO-KNN takes less execution time with used datasets. It is slightly faster than all the evaluated algorithms, especially for Australian dataset.

4.3.3 Results for Research Question 3

In this research question, we aim to explore the statistical robustness of feature selection on the classification accuracy. First, the t-test is performed to test whether the means accuracies of different classifiers are equal. Table 6 presents the performance of the SMPSO-KNN model compared to other tested models. It is observed that the p-value is less than 0.05 for all the obtained results, therefore, the H1 hypothesis is accepted. Consequently, we reject the null hypothesis which means that the performance of the classifiers is equal. The conducted T-test results exhibit that the gain of feature selection differs between all the studied classifiers. To conclude, the efficiency of the feature selection algorithm is basically

Fig. 6 Execution time of proposed classifiers based on feature selection algorithms generated by each dataset

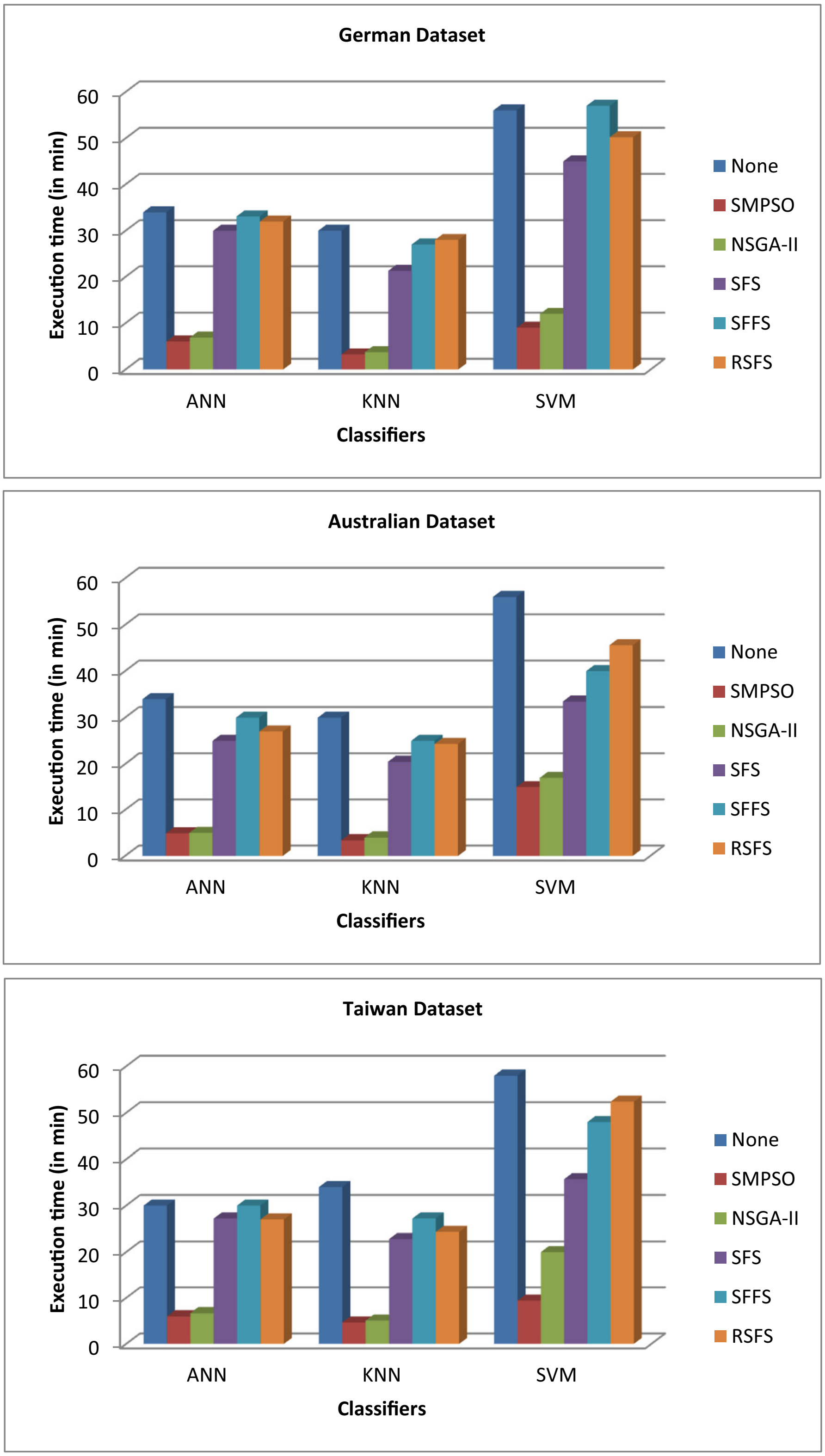


Table 6 Paired t-test for the comparison of the three classifiers performance with FS

Classifiers 1	Classifier 2	t-score	p-value
German credit dataset			
SMPSO-KNN	NSGAII-ANN	14.6920	0.021
	NSGAII-KNN	16.9217	0.010
	NSGAII-SVM	12.9065	0.049
	SFS-ANN	25.1434	0.017
	SFS-KNN	3.0237	0.047
	SFS-SVM	12.6422	0.004
	SFFS-ANN	3.4634	0.037
	SFFS-KNN	4.6457	0.021
	SFFS-SVM	3.4241	0.058
	RSFS-ANN	17.265	0.025
	RSFS-KNN	6.658	0.021
	RSFS-SVM	13.458	0.002
Australian credit dataset			
SMPSO-KNN	NSGAII-ANN	9.251	0.001
	NSGAII-KNN	7.362	0.007
	NSGAII-SVM	13.1548	0.022
	SFS-ANN	4.2150	0.001
	SFS-KNN	3.2541	0.058
	SFS-SVM	4.5412	0.002
	SFFS-ANN	4.5163	0.012
	SFFS-KNN	6.6621	0.025
	SFFS-SVM	4.8741	0.031
	RSFS-ANN	16.9217	0.002
	RSFS-KNN	12.9065	0.004
	RSFS-SVM	25.1434	0.001
Taiwan credit dataset			
SMPSO-KNN	NSGAII-ANN	0.692	0.015
	NSGAII-KNN	16.9217	0.010
	NSGAII-SVM	12.9065	0.049
	SFS-ANN	25.1434	0.017
	SFS-KNN	3.0237	0.047
	SFS-SVM	12.6422	0.044
	SFFS-ANN	3.4634	0.037
	SFFS-KNN	4.6457	0.021
	SFFS-SVM	3.4241	0.048
	RSFS-ANN	17.265	0.025
	RSFS-KNN	6.658	0.021
	RSFS-SVM	13.458	0.002

relying on the classifier used during the feature selection process.

As shown in the previous experiments, SMPSO-KNN provides the optimal subset of features compared to the other applied algorithms in terms of accuracy and F1-score.

Table 7 The statistical significance of each selected features for German dataset

Features	X-squared	df	p-value
7	61.691	34	1.279e-12
3	33.299	9	0.0001185
10	36.099	46	2.761e-07
6	28.368	4 6	0.001045
5	19.1726	9	0.02789
13	57.44	52	0.02795

Tables 7, 8, and 9 summarize the results obtained using the Chi-squared test on the features selected from the used dataset. The results indicate that p-value is always less than the threshold ($p < 0.05$). Basically, the smaller p-value provides significant results. Statistically, these results confirm the strong relationship between the selected feature and the credit risk decision (good and bad).

4.3.4 Result of Research Question 4

In order to predict the credit risk, many works have been conducted to study the impact of feature selection methods on machine learning classifiers. In this experiment, we compare the proposed model with other existing works based on two evaluation metrics accuracy and F1-score. From Table 10, it is observed that the SMPSO-KNN is achieved promising results compared to existing models (Zhang et al., 2018; Jadhav et al., 2018; Tripathi et al., 2018; Sivasankar et al., 2020). The SMPSO-KNN model generates the best result with more than 93% of accuracy followed by IGDFS-SVM (Jadhav et al., 2018) with an average accuracy of 91%. In addition, SMPSO-KNN achieved the highest result for the three datasets regarding the F1-score followed by PS-MLFN model (Tripathi et al., 2018). The conducted experiments confirm the effectiveness of the proposed model for classifying customer credit risk. To this end, we can suggest that our proposed SMPSO-KNN technique is a feasible solution to enhance the performance of the credit risk assessment.

Table 8 The statistical significance of each selected features for Australian dataset

Features	X-squared	df	p-value
2	145.25	29	0.04771
5	239.05	214	0. 01154
4	203.41	22	$< 2.2e - 16$
8	219.3	170	0.006377
3	214.15	131	6.17e-06

Table 9 The statistical significance of each selected features for Taiwan dataset

Features	X-squared	df	p-value
5	30049	4	$< 2.2e - 16$
10	30160	112	$< 2.2e - 16$
3	32343	22	$< 2.2e - 16$
12	51950	44	$< 2.2e - 16$
16	32199	20	$< 2.2e - 16$
22	36190	13	$< 2.2e - 16$

5 Related Works and Discussion

Tripathi et al. (2018) studied the efficiency of feature selection in the credit risk classification. This study includes three steps. The first one is the pre-processing step which aims to analyze the missing data and to transform and normalize dataset. In the second step, a feature clustering-based model is proposed, in which a correlation coefficient method has been applied in order to generate the best clusters of initial features. Finally, the efficiency of feature clustering is proven based on five classifiers (PNN, MLFN, NB, RBFN, and DT). The performance of these different classifiers is determined used weighted voting method which is implemented for aggregating the output of the heterogeneous classifiers. This method aims to ameliorate the predictive performance of proposed credit assessment model. It affects the weight to each used classifiers based on its classification accuracy.

Indeed, (Oreski & Oreski, 2014) argued the importance of features selection for improving classification accuracy. In this study, they combined a hybrid genetic algorithm (HGA) and neural network (HGA-NN) in order to identify the best feature subset and to enhance the accuracy and comprehensibility of credit risk assessment. First, a preliminary step is proposed to reduce the search space. This dimensionality reduction is given by the generation of an initial subset of feature using five filter techniques: earlier experience, information gain, gain ratio, gini index, and correlation. Then, the obtained features subset is used by HGA as the initial population. This reduction makes the search space limited compared to the whole search space. The experimental results show that filtering and reducing the search space improve the performance of the classifier and minimize the computational cost.

Zhang et al. (2018) developed a credit risk assessment model based on an improved version of sequential minimal optimization (SMO) algorithm called four-variable SMO (FV-SMO). Metawa et al. [53] introduces a novel FS model used elephant herd optimization (EHO). The effectiveness of EHO algorithm is evaluated based +on modified water wave optimization algorithm-based deep belief network (MWWO-DBN) classifier. The results confirm that the choice EHO algorithm leads to enhanced classification performance. Jadhav et al. (2018) presented a novel approach for credit classification using a feature selection-based algorithm called Information Gain Directed Feature Selection algorithm (IGDFS). This algorithm sorts the features using the information gain technique in order to generate the top N features important for the decision process. The performance of IGDFS is compared with wrapper genetic algorithm (GAW) based on three classifiers: KNN, Naïve Bayes. Three datasets are used to evaluate the proposed approach. The result shows that IGDFS algorithm commonly outperforms GAW and SVM provide the highest classification accuracy compared to KNN and Naive Bayes algorithm. Thus, the combination of IGDFS and SVM algorithms provide the best result.

Maldonado et al. (2017) proposed a support vector machine based model which is capable to simultaneously classify credit scoring and select the relevant features using an embedded method. This study highlights variable acqui-

Table 10 Accuracy and F1-score comparison

	Method	Accuracy (%)	F1-score (%)
German credit data	SMPSO-KNN	95	94
	FV-SMO (Zhang et al., 2018)	76.80	77.30
	IGDFS-SVM (Jadhav et al., 2018)	82.80	-
	PS-MLFN (Tripathi et al., 2018)	75.28	83.66
	PS-NB (Tripathi et al., 2018)	75.34	82.65
	WABEM (Sivasankar et al., 2020)	83.85	-
Australian credit data	SMPSO-KNN	97	96
	IGDFS-SVM (Jadhav et al., 2018)	90.75	-
	PS-MLFN (Tripathi et al., 2018)	85.62	87.57
	PS-NB (Tripathi et al., 2018)	70.51	78.32
	WABEM (Sivasankar et al., 2020)	89.91	-
Taiwan credit data	SMPSO-KNN	93	96
	IGDFS-SVM (Jadhav et al., 2018)	82.57	-

sition cost taking in consideration multivariate factor and the inter-dependence between features. Two novel Mixed-Integer Programming approaches are presented. However, this study is applied only where features are collected from various sources and at different costs. Verbiest et al. (2016) applied five wrappers training set selection (TSS) where the candidate subset is evaluated using SVM. The result of this study demonstrates that the wrapper approach outperforms the filter one. Sadatrasoul et al. (2015) studied four FS methods in pre-processing step to filter data. The result of this step is combined with the fuzzy apriori algorithm to generate credit assessment rules. This study evaluated four FS techniques: Pearson correlation, PCA, CART, and fuzzy apriori FS. The result shows that fuzzy apriori FS outperforms the other compared techniques.

Sivasankar et al. (2020) confirmed that FS is a crucial step which ameliorates the accuracy of credit risks prediction model. Sivasankar et al. suggested a weight-adjusted boosting ensemble method (WABEM) using rough set based feature selection with the balancing and regression-based preprocessing (RS-RFS-WABEM). A balancing process is used by undersampling the targets classes. Then, a regression step is applied to fill missing values. The conducted experiments show that this proposed ensemble model outperforms other base and ensemble classifiers such as bagging and random subspace ensemble classifiers. Thus, RS-RFS-WABEM model is considered effective model to predict risky applicants in a better way. Arora and Kaur (2020) applied Bolasso algorithm to select subset of relevant and important features from the full subset of features. The selected features will be used then by various classifiers such as SVM, Random Forest, K-NN and Naïve Bayes to evaluate the effectiveness of Bolasso algorithms to determine the relevant features. It is noticed that Bolasso enabled Random Forest Algorithm provides the highest accuracy for credit risk evaluation.

Nevertheless, the majority of the aforementioned studies have proposed different approaches for credit risk classification that use features selection algorithms to discover the most important features in pre-processing phase. In this paper, we studied the efficiency of SMPSO, NSGA-II, SFS, SFFS, RSFS as FS algorithms combined with three popular classifiers (ANN, KNN, and SVM) for credit risk classification. The aim of our work is to study the impact of these FS algorithms to enhance the accuracy of credit risk prediction. Our experiments show that SMPSO as FS combined with KNN classifier provides a promising result with an average of more than 93% of accuracy. In fact, the advantages of performing features selection with KNN lies in potentially minimized effort gathering, and treating attributes at data collection. It is also designed as a regularization method which aims to avoid overfitting. Another advantage of this combination is reduced computation time, as shown in the Section 4.3.2.

6 Conclusion

The profit of banks has been increased by predicting the risky applicants effectively. In this way, credit risk assessment model supports the decision making process for finding the risky applicants. FS is a crucial step which is mainly used to select the important features which enhance the performance of classifiers. In this paper, a SMPSO algorithm is used in order to determine the relevant features that ameliorate the effectiveness of applicant classification process. A comparative study is conducted to study the efficiency of SMPSO, NSGA-II, SFS, SFFS, RSFS as FS algorithms combined with three popular classifiers (ANN, KNN, and SVM) for credit risk classification. The results show that the proposed SMPSO-KNN model search the optimal subset of features which provide the best accuracy. This model tends to search the optimal sub-set of features using considering simultaneously two objectives: 1) minimizing the number of features and 2) maximizing the importance (weight) of the selected features. We conducted, also, a statistical test based on the chi-square test to verify more the relationship between the selected feature and credit risk. All attributes have a p-value less than 0.05 which argue the strong relationship between the selected features and the credit risk and prove the relevance of the obtained subsets. Besides, we attend a comparative with similar works; the result of this comparison proves again that SMPSO-KNN model capable to generate the highest result. From this study, we can conclude that the application of SMPSO in feature selection has demonstrated significant improvements in accuracy and execution time. Researchers can leverage this method to explore high-dimensional datasets effectively while maintaining a balance between model complexity and interpretability. Our findings demonstrate that by implementing the SMPSO-KNN model, financial institutions can achieve higher accuracy in predicting risky applicants, leading to more informed decision-making and potentially increased profitability due to better risk management. Still, financial institutions should ensure proper preprocessing of data, maintain diversity during feature selection to avoid premature convergence, and regularly update the model with new data to keep it relevant. Furthermore, Future studies could explore the integration of SMPSO with other classifiers or its application in different domains such as medical diagnosis, fraud detection, or market analysis. Additionally, investigating the combination of SMPSO with other optimization algorithms could yield interesting results.

Author Contributions Ines Gasmi: Conceptualization, Methodology, Formal analysis, Writing - original draft, Visualization.
Sana Neji: Data curation, Software, Writing - editing.
Salima Smiti: Writing - review, software, validation.
Makram Soui: Supervision, Conceptualization, Methodology, Writing - review.

Funding No funding was received for conducting this study.

Data Availability The datasets used and analyzed during the current study are publicly available from the University of California, Irvine (UCI) Machine Learning Repository.
1. german dataset:
https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.
2. Australian dataset:
https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval
3. Taiwan dataset:
https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
The datasets are used in accordance with the terms and conditions specified by the UCI Machine Learning Repository.

Declarations

Competing Interests None.

References

Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F. (2016). Feature selection using information gain for improved structural-based alert correlation. *PloS One., 11*, 0166017. https://doi.org/10.1371/journal.pone.0166017

Arora, N., & Kaur, P. D. (2020). A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing.,86*,. https://doi.org/10.1016/j.asoc.2019.10593

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review., 45*, 1–23. https://doi.org/10.1007/s10462-015-9434-x

Clerc, M., & Kennedy, J. (2002). The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computation.*,https://doi.org/10.1109/4235.985692

Coello, C.C., Lechuga, M.S. (2002). Mopso: A proposal for multiple objective particle swarm optimization. In: Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600), pp. 1051–1056. https://doi.org/10.1109/CEC.2002.1004388

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning., 20*, 273–297. https://doi.org/10.1007/BF00994018

Doumpos, M., Lemonakis, C., Niklis, D., Zopounidis, C., Doumpos, M., Lemonakis, C., Niklis, D., & Zopounidis, C. (2019). Introduction to credit risk modeling and assessment. *Analytical Techniques in the Assessment of Credit Risk: An Overview of Methodologies and Applications,* 1–21

El Akadi, A., El Ouardighi, A., & Aboutajdine, D. (2008). A powerful feature selection approach based on mutual information. *International Journal of Computer Science and Network Security., 8*, 116.

Ferreira, F. A., Meidutė-Kavaliauskienė, I., Zavadskas, E. K., Jalali, M. S., & Catarino, S. M. (2019). A judgment-based risk assessment framework for consumer loans. *International Journal of Information Technology & Decision Making., 18*(01), 7–33. https://doi.org/10.1142/S021962201850044X

Grabczewski, K., Jankowski, N. (2006). In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.) Mining for Complex Models Comprising Feature Selection and Classification, pp. 471–488. Springer, Berlin, Heidelberg . https://doi.org/10.1007/978-3-540-35488-8_24

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in p2p lending. *European Journal of Operational Research., 249*(2), 417–426. https://doi.org/10.1016/j.ejor.2015.05.050

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research., 3*, 1157–1182.

Hancer, E., Xue, B., Zhang, M., Karaboga, D., & Akay, B. (2018). Pareto front feature selection based on artificial bee colony optimization. *Information Sciences., 422*, 462–479. https://doi.org/10.1016/j.ins.2017.09.028

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). Mifs-nd: A mutual information-based feature selection method. *Expert Systems with Applications., 41*, 6371–6385. https://doi.org/10.1016/j.eswa.2014.04.019

Huang, B., Buckley, B., & Kechadi, T.-M. (2010). Multi-objective feature selection by using nsga-ii for customer churn prediction in telecommunications. *Expert Systems with Applications., 37*, 3638–3646. https://doi.org/10.1016/j.eswa.2009.10.027

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the ga algorithm for feature selection. *Journal of Big Data., 9*(1), 24. https://doi.org/10.1186/s40537-022-00573-8

Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events theoretical background. *Int J Eng Res Appl.,3*, 605–610.

Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing., 69*, 541–553. https://doi.org/10.1016/j.asoc.2018.04.033

Jain, I., Jain, V. K., & Jain, R. (2018). Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing., 62*, 203–215. https://doi.org/10.1016/j.asoc.2017.09.038

Jin, X., Xu, A., Bie, R., Guo, P. (2006). Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In: Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore, April 9, 2006. Proceedings, pp. 106–115 . https://doi.org/10.1007/11691730_11

Kang, B.-Y., Kim, D.-S., & Kang, S.-H. (2010). Extended knn imputation based lof prediction algorithm for real-time business process monitoring method. *The Journal of Society for E-Business Studies., 15*, 303–317. https://doi.org/10.1016/j.asoc.2018.04.033

Kasongo, S. M. (2021). An advanced intrusion detection system for iiot based on ga and tree based algorithms. *IEEE Access., 9*, 113199–113212. https://doi.org/10.1109/ACCESS.2021.3104113

Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications., 76*, 1–11. https://doi.org/10.1016/j.eswa.2017.01.048

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services., 27*, 11–23. https://doi.org/10.1016/j.jretconser.2015.07.003

Kramer, O. (2013). K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors,* 13–23

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications., 40*, 5125–5131. https://doi.org/10.1016/j.eswa.2013.03.019

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)., 50*(6), 1–45. https://doi.org/10.1145/3136625

Mafarja, M., Aljarah, I., Faris, H., Hammouri, A. I., & Ala'M, A.-Z., Mirjalili, S. (2019). Binary grasshopper optimisation algorithm approaches for feature selection problems. *Expert Systems with Applications.,117*, 267–286. https://doi.org/10.1016/j.eswa.2018.09.015

Mafarja, M., & Mirjalili, S. (2018). Whale optimization approaches for wrapper feature selection. *Applied Soft Computing., 62*, 441–453. https://doi.org/10.1016/j.asoc.2017.11.006

Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research., 261*, 656–665. https://doi.org/10.1016/j.ejor.2017.02.037

Marcano-Cedeño, A., Quintanilla, J., Cortina-Januchs, G., Andina, D. (2010). Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: 2010 IEEE 36th Annual Conference on Industrial Electronics Society, pp. 3945–3950. https://doi.org/10.1109/IECON.2010.5675075

Nebro, A.J., Durillo, J.J., Garcia-Nieto, J., Coello Coello, C.A., Luna, F., Alba, E. (2009). Smpso: A new pso-based metaheuristic for multi-objective optimization. In: 2009 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making, pp. 66–73 . https://doi.org/10.1109/MCDM.2009.4938830 . IEEE

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications., 41*, 2052–2064. https://doi.org/10.1016/j.eswa.2013.09.004

Pohjalainen, J., Räsänen, O., & Kadioglu, S. (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language., 29*, 145–171. https://doi.org/10.1016/j.csl.2013.11.004

Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X., & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing., 74*, 634–642. https://doi.org/10.1016/j.asoc.2018.10.036

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research.* 3(Mar), 1371–1382

Rostami, M., Berahmand, K., & Forouzandeh, S. (2021). A novel community detection based genetic algorithm for feature selection. *Journal of Big Data., 8*(1), 2. https://doi.org/10.1186/s40537-020-00398-3

Sadatrasoul, S., Gholamian, M., & Shahanaghi, K. (2015). Combination of feature selection and optimized fuzzy apriori rules: the case of credit scoring. *Technol: Int. Arab J. Inf.,*

Sivasankar, E., Selvi, C., & Mahalakshmi, S. (2020). Rough set-based feature selection for credit risk prediction using weight-adjusted boosting ensemble method. *Soft Comput., 24*, 3975–3988. https://doi.org/10.1007/s00500-019-04167-0

Song, Q., Jiang, H., & Liu, J. (2017). Feature selection based on fda and f-score for multi-class classification. *Expert Systems with Applications., 81*, 22–27. https://doi.org/10.1016/j.eswa.2017.02.049

Soui, M., Gasmi, I., Smiti, S., & Ghédira, K. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. *Expert systems with applications., 126*, 144–157. https://doi.org/10.1016/j.eswa.2019.01.078

Su, C.-T., & Lin, H.-C. (2011). Applying electromagnetism-like mechanism for feature selection. *Information Sciences., 181*, 972–986. https://doi.org/10.1016/j.ins.2010.11.008

Tran, B., Xue, B., & Zhang, M. (2016). Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing., 8*, 3–15. https://doi.org/10.1007/s12293-015-0173-y

Tripathi, D., Edla, D. R., Kuppili, V., Bablani, A., & Dharavath, R. (2018). Credit scoring model based on weighted voting and cluster based feature selection. *Procedia Computer Science., 132*, 22–31. https://doi.org/10.1016/j.procs.2018.05.055

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics., 85*, 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

Verbiest, N., Derrac, J., Cornelis, C., García, S., & Herrera, F. (2016). Evolutionary wrapper approaches for training set selection as preprocessing mechanism for support vector machines: Experimental evaluation and support vector analysis. *Applied Soft Computing., 38*, 10–22. https://doi.org/10.1016/j.asoc.2015.09.006

Wang, J., Neskovic, P., & Cooper, L. N. (2007). Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters., 28*, 207–213. https://doi.org/10.1016/j.patrec.2006.07.002

Wang, D., Zhang, Z., Bai, R., & Mao, Y. (2018). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. *Journal of Computational and Applied Mathematics., 329*, 307–321. https://doi.org/10.1016/j.cam.2017.04.036

Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions on Cybernetics., 43*, 1656–1671. https://doi.org/10.1109/TSMCB.2012.2227469

Yu, L., Liu, H. (2004). Feature selection for high-dimensional data: A fast correlation-based filter solution. Unpublished manuscript

Zhang, Y., Wang, S., Phillips, P., & Ji, G. (2014). Binary pso with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems., 64*, 22–31. https://doi.org/10.1016/j.knosys.2014.03.015

Zhang, Y., Gong, D., Hu, Y., & Zhang, W. (2015). Feature selection algorithm based on bare bones particle swarm optimization. *Neurocomputing., 148*, 150–157. https://doi.org/10.1016/j.neucom.2012.09.049

Zhang, Q., Wang, J., Lu, A., Wang, S., & Ma, J. (2018). An improved smo algorithm for financial credit risk assessment – evidence from china's banking. *Neurocomputing., 272*, 314–325. https://doi.org/10.1016/j.neucom.2017.07.002

Zhu, Y., Liang, J., Chen, J., & Ming, Z. (2017). An improved nsga-iii algorithm for feature selection used in intrusion detection. *Knowledge-Based Systems., 116*, 74–85. https://doi.org/10.1016/j.knosys.2016.10.030

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Ines Gasmi is a computer science educator specializing in credit risk prediction. She earned her master's degree in 2015 from the Higher Institute of Management of Gabes, Tunisia, and completed her Ph.D. in 2020 at the University of Manouba, Tunisia, focusing on enhancing credit risk prediction. She was a teaching assistant at the Higher Institute of Management of Gabes. Her research interests encompass credit scoring, feature selection, and swarm intelligence.

Sana Neji is a dedicated computer science lecture with extensive experience teaching at several Michigan universities. She currently serves as a Lecturer III in Computer Science at the University of Michigan-Flint's College of Innovation and Technology. In this role, she teaches courses such as Database Management Software. Prior to her position at UM-Flint, Sana Neji taught at the University of Michigan-Dearborn and Oakland University, where she instructed courses including Computer Literacy and Information Management, Discrete Structures I, Discrete Structures II, and Software Engineering I.

Salima Smiti obtained her baccalaureate degree in 2011 from Mohamed Ali School Elhamma. She obtained her diploma and her master degree in Computer Sciences from the Higher Institute of Management of Gabes in 2014 and in 2017. Since 2018, she is pursuing her doctoral studies at the National School of Computer Science of Manouba, Tunisia (ENSI), and she is a member of Research Groups on COSMOS research laboratory. She is currently teaching at the Higher Institute of Management of Gabes. Her current research interests are focused on the application of machine learning and deep learning techniques on the field of the financial institution.

Makram Soui received a PhD in Computer Science from the Polytechnic University of Hauts-de-France, in 2010. He holds many academic certificates such as IBM Predictive Analytics Modeler, IBM Business Intelligence Analyst, Oracle Certified Professional Java Programmer, IBM Cloud Application Developer, Microsoft 98-367 Security fundamentals. He is currently an assistant professor in engineering and computer science, university of Michigan-Flint. He teaching many courses such as artificial intelligence, machine learning, datamining, to undergraduate and postgraduate students, revising and developing courses plans. He published 10 referred journal papers and 24 conference papers with a low acceptance rate (between 22% and 34%).