Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **9**, Number **2**, May **2024**

Employment Prediction Models Based on Weighted Feature Selection and Semi-Supervised Machine Learning

Huan-Huan Chen

Department of Computer
Hebei University of Water Resources and Electric Engineering, Cangzhou 061000, P. R. China
gongjiankenan163@163.com

Ya-Juan Li*

Department of Computer
Hebei University of Water Resources and Electric Engineering, Cangzhou 061000, P. R. China
yiqinkongjian@163.com

Cheng Peng

Department of Electrical Engineering
Hebei University of Water Resources and Electric Engineering, Cangzhou 061000, P. R. China
czyalongwan666@163.com

Wei-Jie Lee

College of Languages and Communication
Universiti Pendidikan Sultan Idris, Tanjung Malim 35900, Perak, Malaysia
fp0363@163.com

*Corresponding author: Ya-Juan Li
Received November 12, 2023, revised January 17, 2024, accepted March 21, 2024.

Abstract. *Employment prediction model can predict the employment success rate of college students based on their personal background, academic performance, internship experience and other factors, which helps to understand their competitiveness in the job market. Accurate employment prediction models have high research value for optimising the allocation of educational resources. Therefore, this work proposes an employment prediction model based on weighted feature selection and semi-supervised machine learning. Firstly, to address the problems of high feature dimensions of relevant information, the tendency of discrete between attributes, and the high number of redundant features, the study compares the characteristics and defects of typical feature selection algorithms. Then, in order to accurately characterise classified information with different features and quantitative features, weighted Euclidean distance is introduced on the basis of the feature selection method based on maximum correlation, so as to achieve the purpose of effectively measuring the redundancy. Next, a semi-supervised machine learning method based on XGBoost is introduced to learn and judge the new dataset using a discriminative model form of semi-supervised learning. Finally, the proposed model is objectively evaluated quantitatively by comparing it with other commonly used correlation algorithms used for prediction. The experimental results show that weighted feature selection can effectively select important features, so as to select a subset of features with relatively low feature dimensions and optimal classification performance. The proposed model can make use of a small amount of labelled data in the new dataset to label a large amount of unlabelled data, which further improves the prediction accuracy.*
Keywords: employment prediction; feature selection; feature filtering; semi-supervised learning; XGBoost

1. **Introduction.** Employment prediction models can analyse the employment situation and prospects of different majors and help colleges and universities and educational institutions optimize the allocation of educational resources [1, 2, 3]. By understanding the market demand, schools can adjust their professional settings, offer relevant courses, cultivate talents that better meet the employment market demand, and improve the employment competitiveness of graduates.

Employment prediction models can predict students' employment success based on their personal background, academic performance, internship experience and other factors [4, 5]. This helps universities and students to understand their competitiveness in the job market and take measures in advance to improve their competitiveness, such as increasing internship experience and improving academic performance. The main technical problems to be solved in employment prediction modelling can be classified into feature selection problem and prediction problem [6, 7].

Feature selection occupies an unassailable position in the machine learning process. According to the characteristics of machine learning datasets and different research needs, the applicable feature selection algorithms are different. After decades of development, with the development of the Internet, various kinds of data from the society are becoming more and more numerous and diverse in form [8, 9, 10]. In such an environment, in order to adapt to different situations of data, feature selection algorithms are also gradually developed, from single to diverse, from simple to complex. Most of the feature selection algorithms at the beginning were based on thresholding and had a single approach, and then gradually developed a composite feature selection approach in which multiple feature selection algorithms were used in combination [11]. In terms of supervised or unsupervised, the development from supervised to unsupervised also goes through a period of time. At present, the feature selection algorithm in most cases tends to be based on the correlation between the features, the combination of multiple feature selection methods, and the process is generally unsupervised.

Currently, the focus of this research is mainly on the two main steps for selecting the optimal feature subset. Firstly, according to the actual application scenario, it is crucial to use which search strategy to get the optimal sub-feature set of the dataset; secondly, how to prove that the selected feature set is the optimal feature subset, and in the verification, some affirmations are needed for the evaluation criteria. In summary, most of the current feature selection algorithms perform the selection of the optimal feature subset from the above two perspectives.

The main research objective of this work is to predict whether a person will find a job and what kind of job he/she will find under specific conditions by learning and analysing various relevant features. By exploiting the information in unlabelled data, better classification and prediction can be made, thus improving the effectiveness of employment prediction models.

1.1. **Related Work.** The role of feature selection is to find the best subset of features using the relevance of the category to be distinguished and the redundancy between features.

Venkatesh and Anuradha [12] reviewed various types of methods for feature selection and provided a systematic summary of filtering methods. Filtering methods based on correlation, distance, dependency, approximation, etc. were proposed and the advantages and disadvantages of the filtering methods were discussed. Hoque et al. [13] used a filtering method based on fast clustering and mutual information for feature selection.

Firstly, unsupervised clustering of features is performed quickly using spectral clustering of graphs, and then the mutual information of feature pairs within each cluster is computed to select the subset of features with the largest mutual information. This method greatly reduces the computational complexity of feature selection. However, this method only considers the correlation between features, not the correlation between features and target variables, and the selected feature subset may not retain the optimal predictive power. Solorio-Fernández et al. [14] proposed a hybrid filter-wrapper method, which uses variable cluster analysis for filtering, and then validates it with the wrapper method to achieve unsupervised feature selection. feature selection for supervised learning. This method makes full use of the advantages of filter and wrapper methods. However, this method depends on the effectiveness of the filter method, and if the filter is not effective, the wrapper method will not be able to give full play to its effect. Moreover, this method needs to optimise both the filter and the wrapper, which is a complicated process.

The study of prediction problems is an important direction in the field of machine learning and data mining, and researchers have been exploring and improving prediction methods and algorithms, such as linear regression, decision trees, support vector machines, neural networks, etc [15, 16]. Different models are suitable for different data and problems, and in practice it is necessary to select the appropriate model for prediction. Feature selection is the selection of suitable features for prediction model training, and feature extraction is the automatic extraction of the most important features from raw data using machine learning algorithms. Researchers have not only proposed many methods for feature selection and extraction, but also explored how to combine multiple feature selection or extraction algorithms to further improve prediction performance [17].

Decision tree based machine learning prediction models [18] can be classified into unsupervised, semi-supervised and supervised types. Touati et al. [19] proposed an unsupervised anomaly detection method based on conditional random field optimisation using a decision tree structure to build a normal model for use in different conditions to detect anomalies without labelled data. Castán-Lascorz et al. [20] proposed a supervised prediction method for time series based on integrated regression trees, using bootstrap sampling and linear regression models to obtain more accurate prediction trees and improve prediction performance. However, the feature selection method of the above two methods is simple and the computational complexity is high when the number of features is large. The number of decision trees needs to be pre-set, which is not flexible enough. Decision trees are easily overfitted, which affects the generalisation ability of the model. Compared with decision tree-based prediction algorithms, XGBoost integrates multiple CART regression trees [21, 22], and the final prediction is formed by adding the model and the leaf node output values, which improves the stability and accuracy of the prediction. In addition, XGBoost has built-in regularisations such as L_1 and L_2 , which can control overfitting. Compared with unsupervised and supervised prediction algorithms, semi-supervised learning can significantly reduce the need for annotation and leverage the rich unlabelled data to improve performance, which is more suitable for real-world application scenarios. Semi-supervised learning can be easily applied to online and streaming learning, and the labels can be acquired gradually.

1.2. **Motivation and contribution.** Compared with the wrapper method, the filtering method has two main advantages: (1) it has efficient feature dimensionality reduction performance and is easy to scale; (2) it is independent of a specific classifier. Therefore, the filtering-based feature selection method is chosen in this paper. XGBoost supports missing values and can automatically learn the distribution of missing features, which is

helpful for semi-supervised learning problems. Therefore, in this paper, we try to combine filtering-based feature selection and XGBoost to achieve semi-supervised prediction.

The main innovations and contributions of this work include:

(1) Aiming at the problem that the feature dimension of the relevant information is high, the attributes tend to be discrete from each other, and there are more redundant features, the weighted Euclidean distance is introduced based on the principle of filtering feature selection based on the maximum correlation, and the fluctuation in different data sets and environments is reduced by comprehensively utilising the weighting information of various features.

(2) After completing the weighted feature selection, a semi-supervised machine learning method based on XGBoost is introduced and trained using labelled and unlabelled data to improve the performance and generalisation of the model. A discriminative model form of semi-supervised learning is used to learn and judge the new dataset to further improve the accuracy of prediction.

2. **Weighted feature selection based on maximum correlation.**

2.1. **Role and definition of feature selection.**

Feature selection for student employment prediction is to construct a simple and effective student employment prediction model by automatically selecting a subset of features that contribute significantly to the prediction of student employment outcomes from the information of students' grades, grades, genders, families, etc. by means of an algorithm. It can discover the factors that have the greatest impact on student employment and gain an in-depth understanding of key features.

The role of feature selection is to find the best subset of features using the relevance of the categories to be distinguished and the redundancy between features. Employment data has a large number of redundant features, so feature selection can reduce the number of irrelevant features. Feature selection can automatically identify the most critical subset of features from the complex employment influencing factors, and its role is to build a more concise and efficient employment prediction model, mainly in terms of reducing data dimensions, improving prediction accuracy, increasing model interpretability, balancing the distribution of sample categories, and preventing overfitting, etc. Particularly important is the fact that feature selection can find out the key factors affecting employment in-depth, which is important for building an accurate and interpretable employment prediction model. Especially important is that feature selection can deeply discover the key factors affecting employment, which plays an important role in establishing an accurate and explainable employment prediction model.

2.2. **Filtering based feature selection algorithm.**

Filtering based feature selection algorithms require the calculation of different entropy values for each attribute in different data.

The effect of different features on the data classification result is indicated by the magnitude of entropy value. A large entropy value indicates a great facilitating influence on the classification result, and vice versa, it has little influence on the classification or even affects the accuracy of the classification. Filtering-based feature selection algorithms include feature selection algorithms based on information gain, feature selection algorithms based on mutual information, feature selection algorithms based on chi-square test, and feature selection algorithms based on maximum correlation. The basic flow of the filter-based feature selection method is shown in Figure 1.

(1) Feature selection algorithm based on information gain.

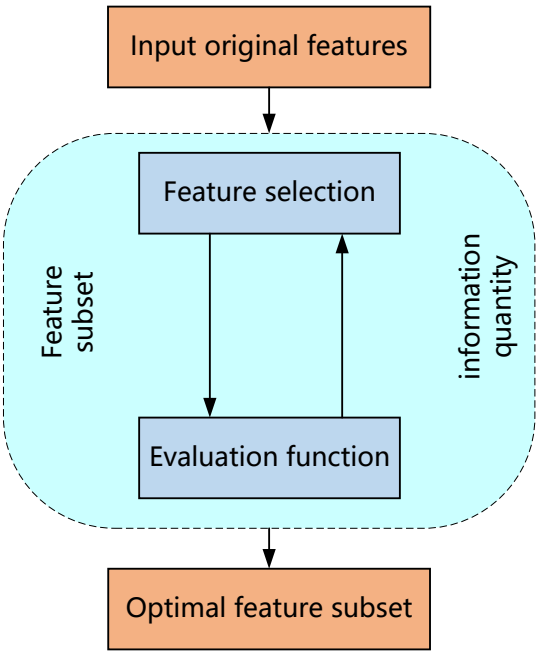

Figure 1. Basic flow of filtered feature selection methods

The feature selection algorithm based on information gain applies information gain to feature selection, so the entropy value of the features is determined by the information gain of each feature [23]. Let the vector $F = (f_1, f_2, \ldots, f_m)$ denote the set of all features and has m sample features. The vector $C = (C_1, C_2, \ldots, C_n)$ denotes the category with n categories to be distinguished. For a given attribute f_i the information gain between the associated category c_i is $\mathrm{IG}(C; F)$.

$$\mathrm{IG}(C; F) = H(C) - H(C \mid F) \tag{1}$$

where $H(C)$ denotes the original entropy value of the system, and $H(C \mid F)$ denotes the average conditional entropy value for C.

Feature selection algorithms based on information gain are able to reduce the dimensionality of the feature set to shrink the feature subset very well and are widely used in the field of intrusion detection. Researchers' experimental results show that the information gain algorithm is still one of the best algorithms for classification today [24]. However, this algorithm still has obvious drawbacks, for example, it only considers the classification impact on the overall target in two scenarios, namely, the occurrence of feature items and the non-occurrence of feature items. If the features are selected inaccurately, then a large amount of noisy data will be generated, which leads to the problem of sparse data.

(2) Mutual information based feature selection algorithm.

The feature selection algorithm based on Mutual Information (MI) applies MI to feature selection, and the measure of this feature selection is to use the size ranking of MI as a way to discriminate the relevance between a feature item and the distinguished category. It reflects the extent to which a feature item is closely related to the category to be discriminated. In general, since the larger MI indicates the closer relation between the two, related studies have selected features with larger MI as the prediction variable of the theory.

MI denotes the amount of information shared between two different variables X and Y, then the mutual information of X and Y can be obtained as:

$$I(x,y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_i)} \tag{2}$$

where $p(x_i, y_j)$ is the joint probability density function of X and Y, while $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y, respectively. So using mutual information as a measure can quantitatively get the correlation between each feature and category, and at the same time the mutual information of both can be calculated in another way.

$$I(f_i, c_i) = \log \frac{p(f_i, c_i)}{p(f_i)p(c_i)} \tag{3}$$

However, in feature selection, the combination of individual feature classification numbers obtained in this way does not completely increase the performance of the classifier, as there is a possibility that the features are highly correlated with each other for reasons thus leading to redundancy in the feature variables. So in this paper, we will propose the scheme to take into account the redundancy in order to select a better subset of features based on it.

(3) Feature selection algorithm based on chi-square test.

The most basic idea of feature selection algorithms based on chi-square test [26] describes the independence of two events by means of chi-square "cardinality", i.e., the degree of influence of a feature on the classification is determined by observing the deviation of the actual value of the two events from the theoretical value as a metric. If the obtained deviation exceeds a certain threshold, it is obvious to the programme that the two are in fact related, and therefore the alternative hypothesis is accepted while the original hypothesis is rejected. The degree of deviation is calculated as shown below:

$$\phi = \sum_{i=1}^{n} \frac{(x - E)^2}{E} \tag{4}$$

where E denotes the theoretical value and x denotes the actual value.

The feature selection algorithm based on the chi-square test, although it is more comprehensive in taking into account the positive and negative correlations between the categories to be distinguished and the individual features, also makes this method more expensive. Therefore, if applied to the field of employment forecasting, it will yield lower forecasting efficiency.

(4) Maximum correlation based feature selection algorithm.

In order to measure the relevance and redundancy of feature items, feature selection algorithms based on maximum correlation utilise MI as a discriminant [27]. A feature is maximally relevant if it has the highest correlation with the category to be distinguished. The purpose of maximum correlation is to select feature items that contain the maximum possible information about the category to be distinguished. However, in order to make the categorical information contained in each feature not overlap, the method introduces minimum redundancy. The purpose of minimum redundancy is that the correlation between the selected feature items is minimised and, at the same time, a subset of features with minimum redundancy is obtained. The maximum correlation and minimum redundancy measures can be obtained by this method.

$$\max D(S,C), D = \frac{\sum\limits_{f_i \int S} MI(f_i, c_i)}{|S|} \tag{5}$$

$$\max R(S), R = \frac{\sum\limits_{f_i, f_j \int S} MI(f_i, f_j)}{|S|^2} \tag{6}$$

where S is the set of final feature selection features, i.e., the selected feature subset, $|S|$ is the number of features contained in the feature subset, C denotes the category to be distinguished, $MI(f_i, c_i)$ is the value of mutual information between a feature item f_i and a category c_i, $MI(f_i, f_j)$ represents the value of mutual information between two different feature items, D denotes the relevance of the feature item to the corresponding category to be distinguished, and R denotes the redundancy between the selected feature items. Redundancy between the selected feature items.

In order to make the relevance of the different selected features to the distinguished categories as high as possible and the redundancy between the selected features as low as possible, the programme summarises the two measures of relevance and redundancy, and the rules obtained are shown below:

$$\max \phi(D, R), \phi = D - R \tag{7}$$

The scheme considers redundancy and relevance, but the relevance and redundancy measurements are mainly based on the amount of mutual information, and this single measure can lead to less than optimal selection of the feature set due to the presence of some features with specific less classification information in the selected feature subset, resulting in less than optimal classification performance of the classifier.

2.3. **Proposed multi-distance weighted feature selection.** In order to improve the efficiency of employment prediction and remove the disturbing data features, this paper firstly adopts the improved maximum correlation method to select the features of student data in order to exclude unnecessary features.

Most dimensionality reduction methods (i.e., methods for filtering features) focus primarily on such features that have the highest correlation with the target class. However, when two features even can have a high degree of interdependence, their contributions to differentiate the target class are still not superimposed. Therefore, based on this this paper proposes a method to measure the independence of each feature based on a multi-distance weighting function. The further the distance, then the higher the independence and the lower the redundancy.

In this paper, Pearson's correlation coefficient is introduced for correlation measurement and weighted Euclidean distance etc. is introduced for redundancy measure. The larger the Pearson's correlation coefficient of different features, the higher the correlation between the feature and the target class. At the same time, if the distance between different features is larger, then it means that the redundancy of the feature subset is lower. Such features having greater correlation and redundancy will be selected for the final feature subset. Finally, the generated feature subset will have low redundancy and strong correlation with the target class.

The chosen feature subset provides the maximum contribution to classification with the target class as the goal, which usually means the minimum error in classification, and the minimum error usually requires the maximum relevance. So this requires that the chosen feature subset is the one that has the highest correlation with the classification target c. In this paper, the Pearson correlation coefficient is introduced to measure the positive and negative correlation between features and the classification target, and also due to its computational efficiency, the Pearson correlation coefficient will be chosen as the correlation measure between features and the target class c in this paper.

Given two vectors $\vec{X}$ and $\vec{Y}$, the Pearson correlation coefficient is calculated as shown below:

$$PC(\vec{X},\vec{Y}) = \frac{S_{XY}}{S_{\vec{X}}S_{\vec{Y}}} \tag{8}$$

$$S_{\mathrm{XY}} = \frac{1}{N-1}\sum_{K=1}^{N}(x_k-\bar{x})(y_k-\bar{y}) \tag{9}$$

$$S_{\vec{x}} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(x_k-\bar{x})^2} \tag{10}$$

$$S_{\vec{y}} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(y_k-\bar{y})^2} \tag{11}$$

$$\bar{x} = \frac{1}{N}\sum_{k=1}^{N}x_k \tag{12}$$

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N}y_k \tag{13}$$

where x_k and y_k are the k-th elements of the vectors $\vec{X}$ and $\vec{Y}$ respectively.

Maximum correlation is then defined:

$$\mathrm{Max\,MR}_i = \left|\mathrm{PCC}(\vec{F}_i,\vec{C}_i)\right| (1 \le i \le M) \tag{14}$$

Where $\vec{F}_i$ denotes a vector consisting of the values of the feature f_i for each data sample, and $\vec{C}_i$ denotes a vector consisting of the values of the classification target to which each data sample belongs.

In many related studies, the m best features can be filtered by the descending order of MR. However, the m best features are not necessarily the best features. Because this method has a defect: "one selected feature may contain the selection information of another feature", i.e., there is redundancy between features and features. Therefore, this paper proposes a method based on weighted Euclidean distance, which can effectively reduce the redundancy between different features.

In a multidimensional data space structure, Euclidean distance is a measure of the distance between two vectors in a multidimensional space. The traditional expression for calculating Euclidean distance is shown below:

$$ED(\overrightarrow{X},\overrightarrow{Y}) = \sqrt{\sum_{k=1}^{N}(X_k-y_k)^2} \tag{15}$$

The traditional Euclidean distance can well represent the cumulative difference between two vectors in space, but it does not take into account the problem of measuring the similarity between the individual feature elements corresponding to the two feature vectors, due to the fact that different elements have different metric information. If the Euclidean distance is used directly between two feature vectors, the two features will make the measure of redundancy inaccurate due to the metric difference.

Therefore, in order to solve the problem of feature vector redundancy measures due to metric differences the computation of weighted Euclidean distance is introduced.

$$NED(\vec{X},\vec{Y}) = \sqrt{\sum_{K=1}^{N} \frac{w_i}{W}(X_k - y_k)^2} \tag{16}$$

$$w_i = e^{-\frac{|x_i - y_i|}{\sigma}} \tag{17}$$

where σ is the adjustment factor, which is used to adjust the weights based on the expertise nuances.

We use maximum distance to measure the similarity between two feature vectors. But in order to combine various dimensions of distance, cosine similarity and Tanimoto coefficient are also introduced in this paper.

$$COS\left(\overrightarrow{X},\overrightarrow{Y}\right) = \frac{\overrightarrow{X}\cdot\overrightarrow{Y}}{\left\|\overrightarrow{X}\right\|\cdot\left\|\overrightarrow{Y}\right\|} \tag{18}$$

$$\left\|\vec{X}\right\| = \sqrt{\sum_{k=1}^{N} x_k^2} \tag{19}$$

$$TC(\overrightarrow{X},\overrightarrow{Y}) = \frac{\overrightarrow{X}\cdot\overrightarrow{Y}}{\left\|\overrightarrow{X}\right\|^2 + \left\|\overrightarrow{Y}\right\|^2 - \overrightarrow{X}\overrightarrow{Y}} \tag{20}$$

Using the above three combined distance-based redundancy measures, we can obtain the Euclidean distance, cosine similarity and Tanimoto coefficient for the i-th feature, in that order:

$$ED_i = \frac{1}{M-1}\sum_{i,j=1}^{M} NED\left(\vec{F_i},\vec{F_k}\right), (1 \le k \le M, K \ne i) \tag{21}$$

$$COS_i = \frac{1}{M-1}\sum_{i,j=1}^{M} COS\left(\vec{F_i},\vec{F_k}\right), (1 \le k \le M, k \ne i) \tag{22}$$

$$TC_i = \frac{1}{M-1}\sum_{i,j=1}^{M} TC(\overrightarrow{F}_i,\overrightarrow{F}_k), (1 \le k \le M, K \ne i) \tag{23}$$

Therefore, the average distance i.e. redundancy can be obtained as:

$$MD_i = \frac{1}{3}\left(ED_i + COS_i + TC_i\right)(1 \le i \le M) \tag{24}$$

After completing the above preparations, we can start selecting the feature subspace. Suppose we have selected the feature subset with $m-1$ features. The next task is to select the m-th feature from the remaining feature set. The algorithm selects the optimal features under the condition:

$$\max(MR_i + MD_i) \tag{25}$$

Calculate the sum of the weighted values of the different features.

$$T = m_1 MR + m_2 MD \tag{26}$$

Selecting the features with larger T will result in the selected important features. In this experiment, the parameters m_2 and m_1 are both equal to 1. Finally, the feature-selected dataset is split into two parts, one part is used for classifier training and the other part is used as a test dataset as an evaluation of the classification performance.

3. Semi-supervised XGBoost-based employment forecasting model.

3.1. Principle of semi-supervised XGBoost.

XGBoost is a decision tree based integration (Ensemble) algorithm [28, 29], the basic idea is to use an additive model with leaf node predictions to integrate multiple decision trees, to reduce the error, as shown in Figure 2.

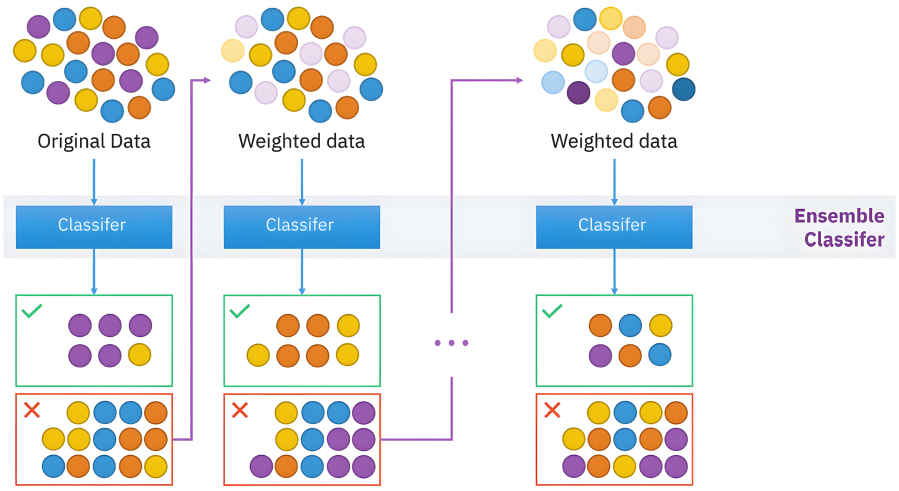

Figure 2. Fundamentals of XGBoost

Semi-supervised XGBoost is based on XGBoost and introduces the use of unlabelled data, the basic idea is to train two XGBoost models simultaneously, one model is used for labelled data and the other model is used for unlabelled data. In each round of Boosting, the labelled data is used for training first, and then the trained model is used to predict the unlabelled data to generate soft labels [30]. Then the soft labels are considered as correct labels and the training is continued by combining labelled and soft labelled data. In this way, the unlabelled data can be used to help the model training. Meanwhile, semi-supervised XGBoost adds a regularity term to minimise the distance between the labelled and unlabelled data in the feature space to make the learned model smoother and more stable.

3.2. The process of implementation of the employment forecasting model.

A tree structure is constructed for the sample set $D = \{(x_i, y_i)\}$ where $|D| = n$, $x_i \in \mathbb{R}^m$, and $y_i \in \mathbb{R}$. The feature vectors x_i and labels y_i are the feature vectors and labels respectively, and the training results of the merged subtrees yield the results shown below:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{27}$$

where $F = \{f(x) = a_{q(x)}\}$ denotes all the subtrees, q is the leaf node, T is the total number of leaves, and the weight occupied by q of each tree f_k is ω. The objective function is established through each subtree as shown below:

$$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{28}$$

$$\Omega(f) = L_1 T + \frac{1}{2} L_2 \|W\|^2 \tag{29}$$

where Ω denotes the regular term, L_1, L_2 are the regular parameters.

In Boosting algorithm, each prediction result is related to the previous prediction result as follows:

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \cdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{cases} \tag{30}$$

where $\hat{y}_i^{(t)}$ denotes the training result of the t-th time, which is obtained by substituting Equation (30) into Equation (29):

$$Obj^{(t)} = L^{(t)} = \sum_{i=1}^{n} l((y_i, y_i^{(t-1)}) + f_i(x_i)) + \Omega(f_t) + cons\tan t \tag{31}$$

Optimising the last item gives:

$$Obj^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{32}$$

The regular term $\Omega(f_t)$ can be transformed as:

$$\Omega(f) = L_1 T + \frac{1}{2} L_2 \sum_{j=1}^{T} w_j^2 \tag{33}$$

Note that L_1 determines whether or not XGBoost's tree continues to fork, while L_2 controls the weight of regularisation. It needs to be set according to the actual situation.

A Taylor expansion of Equation (33) is obtained:

$$Obj = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + L_2) w_j^2] + L_1 T \tag{34}$$

To continue solving Equation (34), we assume the following definition:

$$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i \tag{35}$$

According to Equation (35), Equation (34) is transformed into:

$$Obj^{(t)} = \sum_{j=1}^{M} [G_j \omega_j + \frac{1}{2} (H_j + L2) \omega_j^2] + L_1 M \tag{36}$$

The optimal value of the j-th leaf ω^* and the optimal solution of the objective function obj^* can be obtained according to Equation (36).

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}, obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{37}$$

After finding the most weights of all the leaves, the optimal decision tree structure is obtained.

We perform a descending ordering based on the results obtained in Equation (26) and traverse the feature space using a forward addition strategy. We add one feature at a time, one at a time, and the corresponding feature set is $X_1, X_2, \ldots, X_m$ (where m is the

number of features contained in the feature subset), and then calculate the classification prediction result accuracy of the current feature set by using the semi-supervised XGBoost algorithm, which is obtained as a_i. If $a_i < a_{i-1}$, then the feature x_i is removed from the feature set X, and so on. The cycle continues until the end.

4. Experimental results and analysis.

4.1. Data sources and data preprocessing. In order to verify the performance of weighted feature selection and XGBoost algorithm in the field of employment prediction, Matlab is used for example simulation.

The experimental data comes from the employment management platform of a Chinese university, and the dataset contains the information of graduates for three years from 2019 to 2021. There are more than 86, 208 records in the dataset, of which 70% of the dataset is selected as the training set and the remaining 30% as the test set. Graduate employment data contains data from multiple management statistical system databases, such as school registration management and graduate employment, and it is necessary to make certain integration of the basic school registration information, personal information of graduates, and employment and other relevant features to be analysed and studied in this paper, and carry out data preprocessing of all aspects of the information of the graduates, and select the method of removing or filling the missing values according to the specific situation, and fill or delete the abnormal values. Fill or delete the abnormal values.

Firstly, the research and analysis of the factors influencing the employment of graduates were initially formulated to obtain the following information related to graduates: students' individualised data, their performance in school (including comprehensive and professional grades, extracurricular practices, etc.) and their employment status, of which the basic data of students is shown in Table 1.

Table 1. Student basic information.

Causality	Nicknames	Data type
Name and surname	XM	Varchar(20)
Student number	XH	Varchar(20)
Distinguishing between the sexes	XB	Varchar(2)
(a person's) age	NL	Varchar(20)
Weight	TZ	Varchar(20)
Birthplace	CSD	Varchar (30)
Employment or not	SFJY	Varchar(2)
Name of unit	DWMC	Varchar (50)
Unit address	DWDZ	Varchar (50)
Job Information	GWXX	Varchar (50)
Computer level	JHI	Varchar(20)
...	...	...
English level	YYSP	Varchar(20)

The optimal subset of features derived from the weighted feature selection proposed in this paper is 1-14, totalling 14 features. Then, based on these 14 features, subsequent employment prediction experiments are conducted.

4.2. Comparison of feature selection results. In order to verify the effectiveness of the proposed weighted feature selection algorithm, it is compared with maximum correlation, chi-square test and MI.

Classification predictions were made using the semi-supervised XGBoost algorithm on a subset of features selected by the four feature selection algorithms. Subsequently, the classification accuracy of the classification model is calculated using cross-validation method. The experimental results are examined to make a judgement on the performance of the algorithms proposed in this paper.

The four algorithms mentioned above were used for feature selection on the dataset and trained using semi-supervised XGBoost. The comparison of the results after performing five cross validations is shown in Table 2, where F_{num} denotes the number of features.

Table 2. Construction of prediction model based on different feature selection algorithms

Feature selection algorithm	Fnum	Precision	Recall	F-value
MI	9	0.9468±0.0332	0.945	0.9459
Chi-square (math.) Test	35	0.9654±0.0203	0.963	0.9642
Maximum relevance	16	0.9703±0.0207	0.968	0.9691
Weighted feature selection	14	0.9732±0.0200	0.970	0.9716

It can be seen that the weighted feature selection method proposed in this paper has a Precision of 97.324% and a Recall of 97.0%, both of which are a little better than the other three feature selection methods. The MI-based feature selection algorithm performs better in dimensionality reduction, but the classification accuracy is lower. The feature selection algorithm based on chi-square test performs lower than the other methods in feature dimensionality reduction. The comparison shows that the weighted feature selection method proposed in this paper is able to select a subset of features with relatively low feature dimensionality and optimal classification performance.

4.3. **Effect of regularisation parameter on prediction.** Figure 3 demonstrates the accuracy of employment prediction for different values of L_1 and L_2 for the semi-supervised XGBoost model. It can be seen that the values of L_1 and L_2 have a more significant effect on the accuracy of employment prediction. When $L_1 = 0.2$ and $L_2 = 1$, the accuracy is at its highest point. When $L_1 = 0.83$ and $L_2 = 1$, the accuracy is at the lowest point. Therefore, the regularisation parameters $L_1 = 0.2$ and $L_2 = 1$ are chosen for the XGBoost algorithm in this paper.

4.4. **Comparison of Prediction Accuracy of Different Algorithms.** The training and testing sample sizes are selected as 2000 and 1200 respectively. using Matlab software, the prediction simulation of Random Forest [31], Integrated Regression Tree [20], Gradient Boosting [32] and this paper's algorithm are carried out respectively, and the results are shown in Figure 4.

It can be seen that the employment prediction accuracy improves with increasing prediction time. The Random Forest is the first to start converging when the computing time reaches about 66s. The integrated regression tree begins to converge at about 72s. The prediction accuracy of Gradient Boosting and the algorithm in this paper stabilises at about 78 s. The employment prediction accuracy of the algorithm in this paper is the highest when all the algorithms are stable. When all the algorithms were stable, this paper's algorithm had the highest employment prediction accuracy of 0.93.

4.5. **Prediction time performance.** The prediction time performance of the above four decision tree based algorithms is compared below and the simulation results are shown in Figure 5.

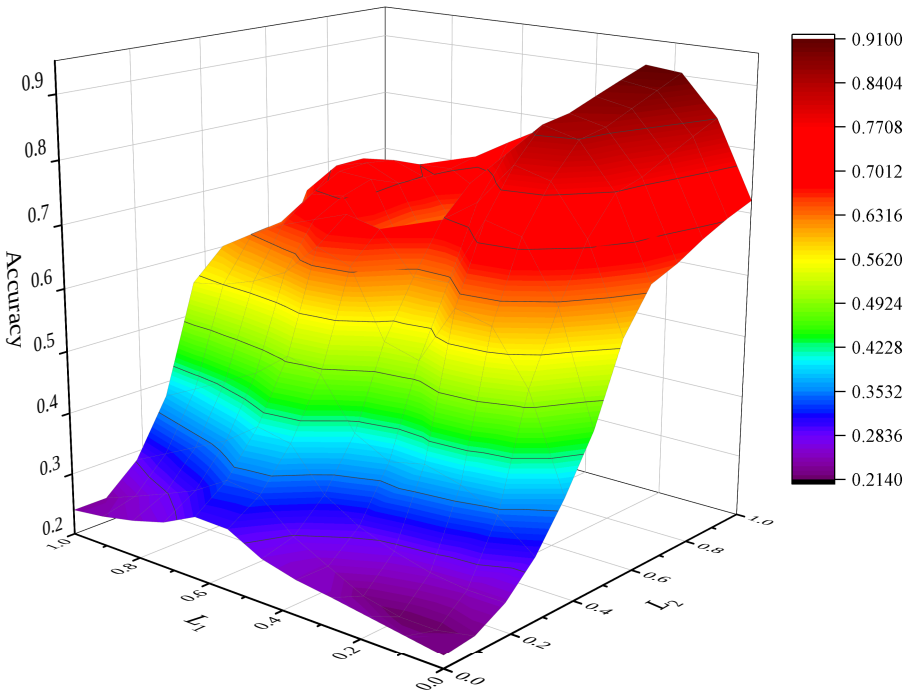

Figure 3. Prediction accuracy with different regularisation parameters

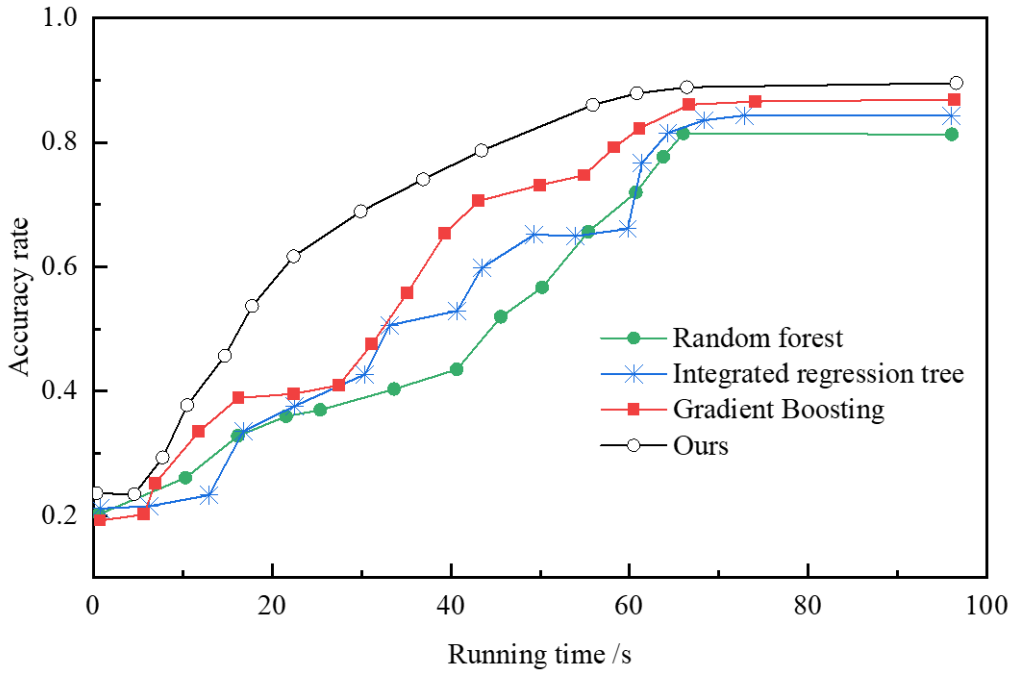

Figure 4. Prediction accuracy of different algorithms

It can be seen that the Random Forest algorithm performs the best and the Integrated Regression Tree algorithm is the worst in terms of prediction time performance. Both Gradient Boosting and the proposed model need to build a tree structure and iterate several times to find the optimal solution, so they consume more time. When the number of student samples to be predicted is large, the proposed model and Gradient Boosting algorithm consume very similar time for employment prediction.

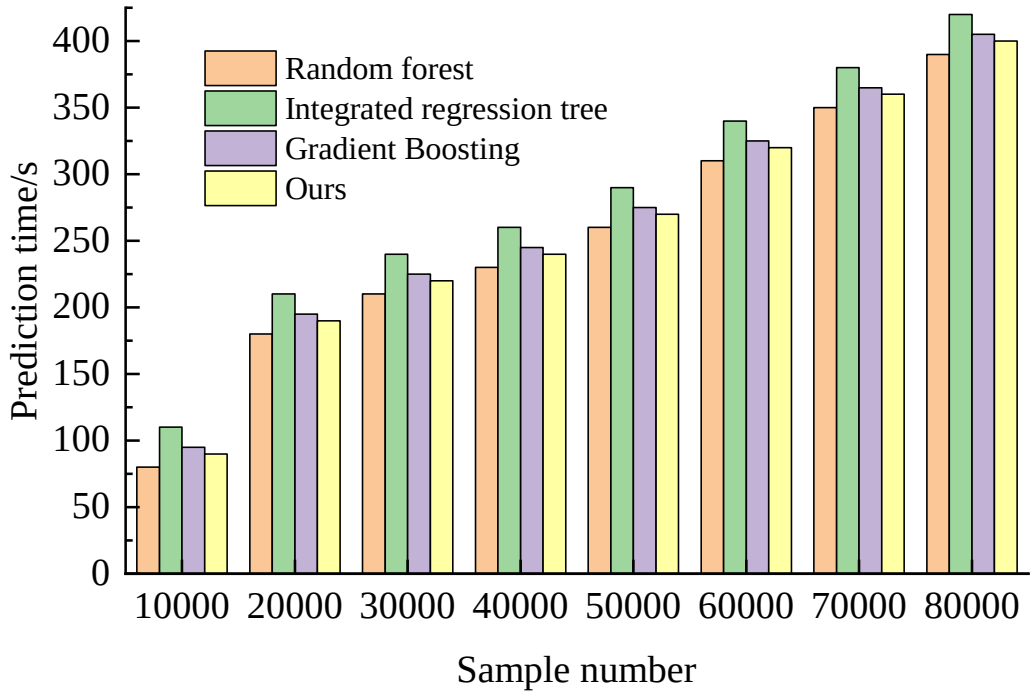

Figure 5. Prediction time of different algorithms

In summary, the proposed model has the highest prediction accuracy and the Random Forest algorithm has the best prediction time performance in terms of employment prediction. Although, the time required by the algorithm in this paper increases when the number of student samples to be predicted is large. However, collectively, the proposed model ensures both high accuracy and no significant increase in prediction time as in the case of the integrated regression tree algorithm.

5. **Conclusion.** This work introduces weighted Euclidean distance based on maximum correlation according to the principle of filtering feature selection, which reduces fluctuations in different datasets and environments by combining the weighting information of various features. After completing the weighted feature selection, a semi-supervised machine learning method based on XGBoost is introduced. The discriminative model form of semi-supervised learning is used to learn and judge the new dataset, which further improves the accuracy of prediction. The comparison shows that the weighted feature selection method proposed in this paper is able to select a subset of features with relatively low feature dimensions and optimal classification performance. The proposed model can make use of a small amount of labelled data in a new dataset to label a large amount of unlabelled data, which improves the prediction accuracy. However, the training time of XGBoost may be long for large-scale datasets. Follow-up studies will try to improve the training speed by tuning the parameters and using distributed computing frameworks such as Dask or Spark.

REFERENCES

[1] Y. A. Chen, R. Li, and L. S. Hagedorn, "Undergraduate international student enrollment forecasting model: An application of time series analysis," *Journal of International Students*, vol. 9, no. 1, pp. 242-261, 2019.

[2] H. Mishchuk, I. Roshchyk, J. Sułkowska, and S. Vojtovič, "Prospects of assessing the impact of external student migration on restoring country's intellectual potential (the case study of Ukraine)," *Economics & Sociology*, vol. 12, no. 3, pp. 209-219, 2019.

[3] A. Malik, E. M. Onyema, S. Dalal, U. K. Lilhore, D. Anand, A. Sharma, and S. Simaiya, "Forecasting students' adaptability in online entrepreneurship education using modified ensemble machine learning model," *Array*, vol. 19, 100303, 2023.
[4] A. Namoun, and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, 237, 2020.
[5] A. Lin, Q. Wu, A. A. Heidari, Y. Xu, H. Chen, W. Geng, and C. Li, "Predicting intentions of students for master programs using a chaos-induced sine cosine-based fuzzy K-nearest neighbor classifier," *IEEE Access*, vol. 7, pp. 67235-67248, 2019.
[6] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.
[7] F. Zhang, T.-Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, "Application of Quantum Genetic Optimization of LVQ Neural Network in Smart City Traffic Network Prediction," *IEEE Access*, vol. 8, pp. 104555-104564, 2020.
[8] F. Zhang, T.-Y. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 2019, 58, 2019.
[9] M.-E. Wu, J.-H. Syu, and C.-M. Chen, "Kelly-Based Options Trading Strategies on Settlement Date via Supervised Learning Algorithms," *Computational Economics*, vol. 59, no. 4, pp. 1627-1644, 2022.
[10] M.-E. Wu, H.-H. Tsai, W.-H. Chung, and C.-M. Chen, "Analysis of Kelly betting on finite repeated games," *Applied Mathematics and Computation*, vol. 373, 125028, 2020.
[11] U. M. Khaire, and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060-1073, 2022.
[12] B. Venkatesh, and J. Anuradha, "A review of feature selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3-26, 2019.
[13] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371-6385, 2014.
[14] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A new hybrid filter–wrapper feature selection method for clustering based on ranking," *Neurocomputing*, vol. 214, pp. 866-880, 2016.
[15] D. J. Putka, A. S. Beatty, and M. C. Reeder, "Modern prediction methods: New perspectives on a common problem," *Organizational Research Methods*, vol. 21, no. 3, pp. 689-732, 2018.
[16] F. Dharma, S. Shabrina, A. Noviana, M. Tahir, N. Hendrastuty, and W. Wahyono, "Prediction of Indonesian inflation rate using regression model based on genetic algorithms," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 45-52, 2020.
[17] Y. Wang, and L. Feng, "A new hybrid feature selection based on multi-filter weights and multi-feature weights," *Applied Intelligence*, vol. 49, pp. 4033-4057, 2019.
[18] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, "Building an efficient intrusion detection system based on feature selection and ensemble classifier," *Computer Networks*, vol. 174, 107247, 2020.
[19] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Transactions on Image Processing*, vol. 29, pp. 757-767, 2019.
[20] M. Castán-Lascorz, P. Jiménez-Herrera, A. Troncoso, and G. Asencio-Cortés, "A new hybrid method for predicting univariate and multivariate time series based on pattern forecasting," *Information Sciences*, vol. 586, pp. 611-627, 2022.
[21] A. Ogunleye, and Q.-G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131-2140, 2019.
[22] O. Sagi, and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Information Sciences*, vol. 572, pp. 522-542, 2021.
[23] B. Azhagusundari, and A. S. Thanamani, "Feature selection based on information gain," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 2, pp. 18-21, 2013.
[24] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowledge-Based Systems*, vol. 54, pp. 298-309, 2013.
[25] X.-F. Song, Y. Zhang, D.-W. Gong, and X.-Y. Sun, "Feature selection using bare-bones particle swarm optimization with mutual information," *Pattern Recognition*, vol. 112, 107804, 2021.
[26] I. S. Thaseen, C. A. Kumar, and A. Ahmad, "Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers," *Arabian Journal for Science and Engineering*, vol. 44, pp. 3357-3368, 2019.

[27] X. Yan, and M. Jia, "Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection," *Knowledge-Based Systems*, vol. 163, pp. 450-471, 2019.

[28] M. W. Dunham, A. Malcolm, and J. K. Welford, "Improved well log classification using semisupervised Gaussian mixture models and a new hyper-parameter selection strategy," *Computers & Geosciences*, vol. 140, 104501, 2020.

[29] W. Liu, K. Deng, X. Zhang, Y. Cheng, Z. Zheng, F. Jiang, and J. Peng, "A semi-supervised tri-catboost method for driving style recognition," *Symmetry*, vol. 12, no. 3, 336, 2020.

[30] R. U. Shaik, A. Unni, and W. Zeng, "Quantum Based Pseudo-Labelling for Hyperspectral Imagery: A Simple and Efficient Semi-Supervised Learning Method for Machine Learning Classifiers," *Remote Sensing*, vol. 14, no. 22, pp. 5774, 2022.

[31] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134, pp. 93-101, 2019.

[32] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937-1967, 2021.