Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 9(90)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.875340

Talent management by predicting employee attrition using enhanced weighted forest optimization algorithm with improved random forest classifier

S. Porkodi[1]*, S. Srihari[2] and N. Vijayakumar[3]
Department of Business Studies, University of Technology and Applied Sciences (HCT), Muscat, Sultanate of Oman[1]
Department of Computer Science, Illinois Institute of Technology, Chicago, USA[2]
IT Instructor, Technical Administrative Training Institute, Muscat, Sultanate of Oman[3]

Received: 19-December-2021; Revised: 05-May-2022; Accepted: 07-May-2022

Abstract

Predictive analysis has been an important field of research suitable for a wide range of applications covering a huge volume of domains in predicting the future with the current and past data. In an organisation, the predicted insights are highly helpful in analysing all the aspects of an issue and making decisions suitably. More specifically, talent management requires making an appropriate decision in employing and maintaining suitable skills in the appropriate place. Machine learning algorithms are most commonly used in analysing the attributes that affect employee attrition and predicting employee turnover. This paper presents the prediction model that makes use of an enhanced weight-based forest optimization algorithm. It employs mutual information for selecting the significant features and a modified random forest for classifying the attrition results. The experimental analysis has been performed with the International Business Machines (IBM) human resource employee attrition dataset and the results are compared with the other existing models. The analysis shows that the proposed model offers better results with an accuracy of 91.23% and a minimum error rate of 8.77% than several other models. The feature significance helps in making effective steps in retaining the talents for the benefit of the organization.

Keywords

Talent management, Employee attrition, Predictive analytics, Machine learning, Random forest, Forest optimization algorithm.

1.Introduction

Talent management and its strategies were getting evolved in the past decades and in recent years, it has been changed at a fast pace [1]. Unlike past decades, talent management policies are becoming indispensable in today’s changing environment [2]. Human resource (HR) is one of the most valuable resources which are to be taken care of in a conscience manner. In general, organizations identify the best talent for the right position and try to engage and retain them effectively to adjust to ever changing conditions [3]. Thus, managing talent is a continuous process that applies a strategic approach to hire, retain, manage as well as develop the workforce [4]. Meanwhile, enterprises must shape a viable workforce by giving proper training insisting on continuous learning and skill development thereby elevating their performance.

Managing talent will be successful only when strategies allow the talents to evolve as the company grows. Human capital management has become more crucial due to the shift in the fundamental nature of today's working environment and the development of technologies [5]. The main challenge for HR professionals is not only to make use of the knowledge the talent has but also to retain them appropriately for the company's benefit. Employee attrition is a workforce in the enterprise who willingly leaves the company [6]. Employee turnover is such a criterion to evaluate the HR professionals that specifies the number of new workers replacing the existing employees over a particular time frame

*Author for correspondence

[7]. Obviously, the higher the employee attrition, the higher the turnover rate cursing huge expenditure on talents for identifying the new hands, providing them training and development to build a competitive nature among them. Proper strategies and policies are to build in the company to retain the talents and to reduce the employee turnover rate. Thus, to predict the possibility of the talent leaving the organization, predictive analysis comes into existence [8].

Data mining and machine learning combined with optimization algorithms has emerged as new age qualitative and quantitative models that are often used to identify clear insights from a wide range of dataset [9]. The predictive analysis makes use of these algorithms for processing the information in hand to predict future trends. Thus, owing to the advancements in the technological field, applying a predictive analysis of the employee attrition data will undeniably help to predict the employees who may leave the organization. This prediction helps to make decisions earlier and to reduce the loss of HR [10, 11].

Classification is the major field of research for many applications covering a wide range of domains. In general, learning algorithms build the model and train them using the training dataset that contains both the features and the target class. The testing phase is intended to apply the model on the test data containing the feature values for predicting the target class [12]. In classification and prediction, the accuracy of prediction is the most important as it allows us to know the level of trust that can keep in the algorithm in connection with the obtained results.
Feature selection is part of the classification process that helps to identify the most important features regarding the target variable and thus it improves the classification performance as well [13]. Using optimization algorithms for soft computing with machine learning for feature selection and classification are more commonly used recently to identify optimal solutions. Most familiarly used optimization algorithms and its variations are ant colony optimization (ACO) [14], a particle swarm optimization (PSO) [15], grey wolf optimization (GWO) [16], forest optimization algorithm (FOA) [17], world cup optimization (WCO) [18], genetic algorithm (GA) [19], animal migration optimization (AMO) [20] and more.

In employee attrition prediction, several researchers applied various standard classification algorithms from machine learning. Each model may provide different results based on its suitability and for employee attrition prediction, it is suggested that the Random forest (RF) algorithm is more appropriate [21, 22]. However, there is still an option for the improvement in the performance of the prediction model. Additionally, most of the works focus only on the classification algorithms and not on the feature selection phase [9–11, 21]. Very few works focus on using optimization algorithms in classifying problems related to HR [15, 22].

Thus, the paper presents the novel prediction model with two phases, including a weighted forest optimization algorithm for feature selection enhanced by utilizing accuracy, specificity, sensitivity along with mutual information and an improved random forest (IRF) algorithm for effectively classifying the attrition results. The experimental analysis for the proposed model has been made using International Business Machines (IBM) human resource employee attrition dataset available for public access and the results of the model are evaluated and compared with the existing models which show improved performance.

The organization of the paper is as follows. Section 2 presents the related research work associated with the field of study. The proposed predictive model for employee attrition with feature selection using an enhanced weighted forest optimization algorithm and IRF classifier are explained in section 3. Section 4 and 5 presents the exploratory data analysis, performance analysis and comparison of results with the existing models. Finally, the paper concludes the article along with the suggestions for future work.

2.Literature review

Talent management is becoming one of the most crucial fields to be taken care of while running an organization. Among all resources, human resource plays a vital role in a company's growth. Retaining talent is such a challenging task since it completely affects the substantial work that contributes to the turnover of the company, if an employee working for the long term at the top-level leaves the enterprise. Due to the consequence of employee attrition, several researchers have used various machine learning algorithms for predicting employee attrition earlier which helps to make strategic decisions without affecting the company's revenue.

To classify the employees based on attrition and retention, extreme gradient boosting (XGBoost), a machine learning technique that uses bagging was implemented [23]. It was claimed that the algorithm has better performance in terms of memory utilization, minimum execution time and maximum accuracy of around 90%. However, the model was not compared with other existing algorithms. Various machine learning classifiers such as decision tree (DT) classifier, logistic regression (LR), support vector machine (SVM) and RF are evaluated with the employee attrition dataset in which the experimental analysis shows that the RF classifier is highly suitable for prediction than other models [7, 20, 24]. It was reported that employee attrition or retention highly depends on their satisfaction level.

Supervised learning models were used to evaluate the employee attrition was made [6, 22]. Various classifiers such as DT, RF, LR, SVM, gradient boosting trees (GBT), XGBoost, neural networks (NN), linear discriminant analysis (LDA), and naive Bayes (NB) and k-nearest neighbour (KNN) were also used. The model was evaluated using varied sizes of small, medium and large datasets. Like a RF classifier, LR is also considered to be effective in predicting employee attrition with an accuracy of 80% [25].

Classification of employee attrition has been made in which C5.0, a DT classifier with Apriori association rule algorithm was utilized which consumes minimum RAM and time consumption [26]. The work was extended with GWO, C5.0 with an association algorithm for employee attrition prediction and the results produced by the algorithm have minimum memory utilization and time consumption than the PSO algorithm. However, the models were implemented with only 5 significant features such as gender, distance from home, environment satisfaction, work-life balance and education field [27]. An ensemble-based model that employs SVM, RF and LR are using weighted average-based classification was suggested to predict the voluntary turnover of the employees. Though the model offers better accuracy, the sensitivity of the model still needs to be improved [9]. Various ensemble models were proposed to improve the prediction accuracy which helps to apply retaining tactics for identified employees [28–31]. Various factors that influence employee attrition along with the possible solution were also discussed.

An algorithm to implement an uplift model that is based on the RF was proposed which helps to evaluate the efficiency of the various retention policies for human resource data [32]. A performance analysis was made between the uplift model and the conventional predictive model and it is shown that the uplift model offers better results in predicting employee turnover [33]. An effort was made for workplace related features that are categorized as demographic and behavioural variables [34]. Another work was made by suggesting the salary increase to proliferate the rate of employee retention via an analytical approach [35].

Thus, with the knowledge obtained from the above literature, it is found that only a very few models focus on selecting the features for the classification process. Feature selection most likely improves the accuracy of the classifiers as it removes the redundant and irrelevant attributes. With this into consideration, a predictive model for talent management has been proposed to effectively classify employee attrition using an enhanced weighted forest optimization algorithm with an IRF classifier.

3.Proposed predictive model

The proposed prediction model aims at managing the talents by predicting employee attrition with a significant feature selection using soft computing and machine learning algorithms. The selection of significant features plays an important role as it represents the key factors that cause the employee to leave the organization and the prediction identifies the experienced employees who have the possibility of leaving the organization based on which proper care can be taken to retain the highly talented professionals. The architecture of the proposed prediction model with enhanced weighted forest optimization and RF algorithms is shown in *Figure 1*. It comprises two phases with the first phase centers on selecting features using a weighted forest optimization algorithm and the second phase on an IRF classifier.

3.1Feature selection using enhanced weighted forest optimization algorithm (FSWFOA)

As mentioned earlier, the feature selection process is more significant as it contributes to the accuracy of classification results for the underlying problem. In general, most real-world datasets consist of an enormous number of features in which all of which may not be useful in the classification or prediction process. This is due to the existence of redundant and irrelevant features occurring in the dataset. Utilizing all such features in any classification or prediction problems may result in inconsistent and inaccurate results. Thus, feature selection identifies the features that are highly dependent on the class variables that even improve the classification accuracy. The FOA is one of the powerful evolutionary algorithms, initially proposed to solve the continuous search space problem. This optimization algorithm is inspired by the procedure of a growth process of limited trees in a forest [36].

However, later it is also adjusted in such a way to use them in discrete search space problems like feature selection [37]. This algorithm is used to identify the optimal feature subset of the entire feature set. Accordingly, the feature selection using the FOA (FSFOA) is divided into six main parts 1) Initialize trees; 2) Local seeding operation (LSO); 3) Population limiting in the forest; 4) Global seeding operation (GSO), and 5) Update the best tree; 6) Fitness function evaluation. The proposed weighted forest optimization algorithm for feature selection enhances the first and last part of the traditional FSFOA algorithm. The FOA algorithm has five main parameters that need to be initialized before executing the algorithms.

They are:

- *Local seeding changes* or (*LSC)* represent the number of features whose values are to be changed on the local seedling stage.
- *Area limit* represents the limitation of the forest indicating the maximum number of trees allowed in the forest.
- *Lifetime* represents the maximum allowed age of the tree.
- *Transfer rate* represents the percentage of the candidate population (cp) to be used in the global seeding stage.
- *Global seeding changes* or (*GSC)* represent the number of the features whose values will be changed in the global seeding stage.

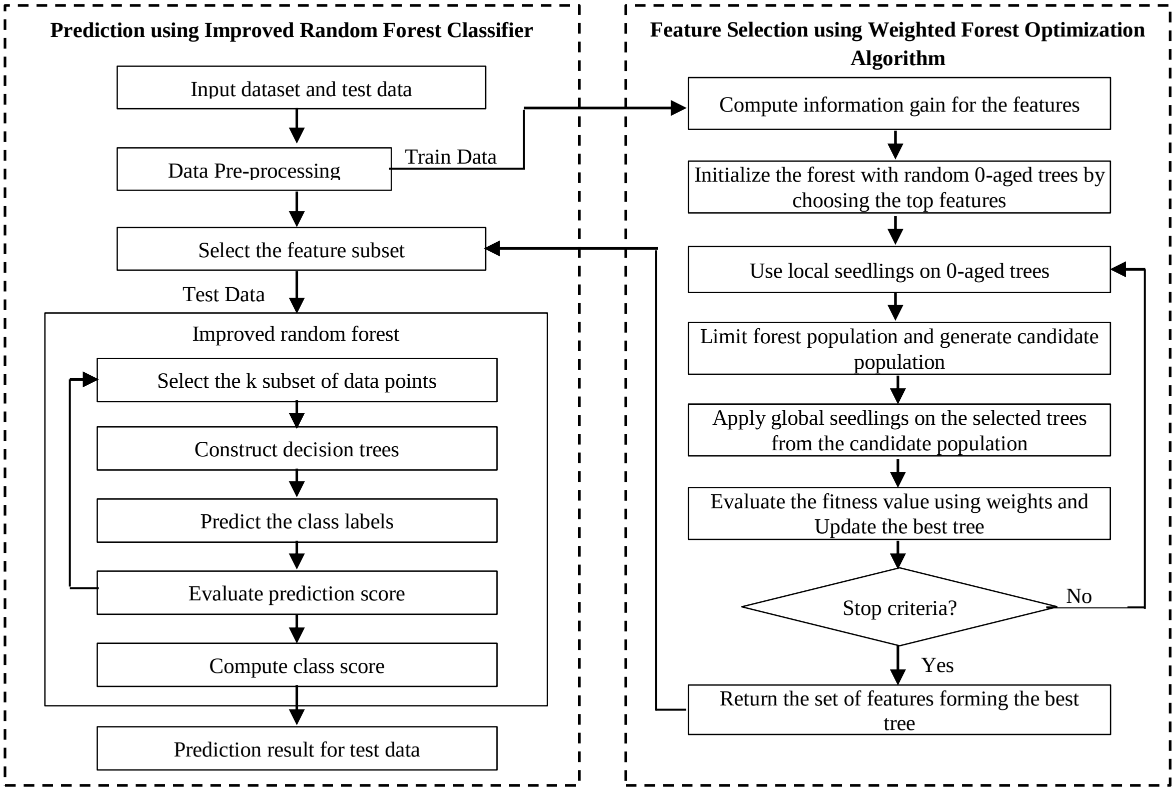

Prediction using Improved Random Forest Classifier
Input dataset and test data
Data Pre-processing
Train Data
Select the feature subset
Test Data
Improved random forest
Select the k subset of data points
Construct decision trees
Predict the class labels
Evaluate prediction score
Compute class score
Prediction result for test data
Feature Selection using Weighted Forest Optimization Algorithm
Compute information gain for the features
Initialize the forest with random 0-aged trees by choosing the top features
Use local seedlings on 0-aged trees
Limit forest population and generate candidate population
Apply global seedlings on the selected trees from the candidate population
Evaluate the fitness value using weights and Update the best tree
Stop criteria?
No
Yes
Return the set of features forming the best tree

Figure 1 The working mechanism of the proposed model

The steps in the FSWFOA algorithm are described in the below subsections.

3.1.1Initialize trees using information gain

In the first stage, the forest is initialized with a randomly generated tree consisting of all sets of features in which every feature is either set as 0 or 1 randomly. Here the size of the tree will be the (n+1) where n represents the number of features in the given dataset and 1 represents the age of the tree which is then initialized to 0. The age of the tree is incremented by 1 in the next stage of local seeding except for the newly generated tree in that stage.

Also, the feature that is set as 1 indicates the feature is selected and 0 indicates the features that are not selected for further learning processing. The sample tree elements in the array representation are given in Equation 1.

$$Tree\ = [age, f_1, f_2, f_3, \dots] \tag{1}$$

However, the main drawback is the completely randomized initial tree generation and it falls into a locally optimal solution. Thus, in the proposed algorithm, initially, information gain (IG) is evaluated for all the *n* features and then the tree is generated by setting the top *n/2* of the features having maximum IG as 1 and for the remaining features, the value is set as 0. The sample initial tree with 4 features in which features f_2 and f_4 are initialized to 1 as they have the highest IG when compared with the other features f_1 and f_3 as in Equation 2.

$$Tree\ = [\,age = 0, f_1 = 0, f_2 = 1, f_3 = 0, f_4 = 1] \tag{2}$$

IG is the most significant concept in DT learning that specifies the ratio of IG to intrinsic information. It specifies the amount of information provided by the feature about a specific class. The IG utilizes entropy values to measure the gain of an attribute [38]. Here the entropy of dataset D is computed as in Equation 3.

$$E(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{3}$$

Here p_i is the probability of the record that belongs to class C_i which can be estimated as $|C_{i,D}|/|D|$. Then the entropy of the attribute A given the information about D is computed as in Equation 4.

$$E_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times E(D_j) \tag{4}$$

Here A is an attribute with *v* distinct values {a_1, a_2, … a_v} in which D is split into *v* partitions and D_j represents the number of records in D having the value of a_j of A.

Finally, the gain of the attribute can be measured as in Equation 5.

$$Gain(A) = E(D) - E_A(D) \tag{5}$$

Finally, the attributes are sorted based on the gain values in descending order which specifies the significance of the attribute concerning the class variable and is used for initializing the features in the initial tree.

3.1.2Local seeding operation

The LSO step is intended to generate a random number of neighbour trees for each tree in the forest in which the *age* of the newly generated tree is set as 0. For the newly generated trees, some features are selected randomly. This is carried out by changing their values from 0 to 1 and from 1 to 0 one at a time. This implies that some features are added or some existing features are removed one at a time before applying the learning algorithm and are considered as the local search in the entire search space. The number of features to be added or removed is the same as the value stored in the predefined parameter *LSC*. Once the local seeding is completed, the *age* of existing trees in the forest is incremented by one except the newly generated trees. The sample local seeding for the tree with four features specified in Equation 2 with LSC as 2 is shown in *Figure 2*. Since LSC is set as 2, the two features f_1 and f_4 are chosen randomly and are processed one at a time in which for the first generated neighbour tree, the feature f_1 is selected along with the features f_2 and f_4 and for the second generated neighbour tree, the selected feature f_4 is removed by selecting only the feature f_2.

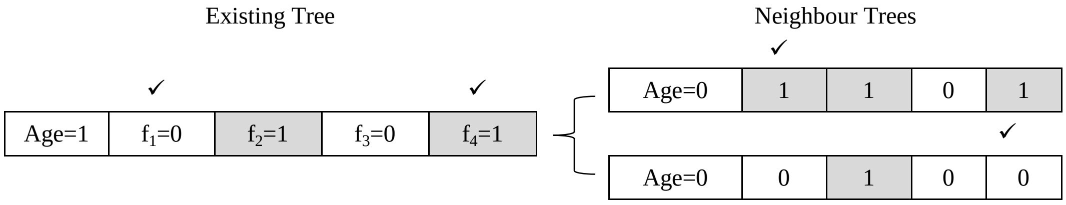


Figure 2 Simple illustration for LSO

3.1.3Limiting the population in the forest

This step is significant as it limits forest expansion. The limitation of the population in the forest can be done by displacing the set of trees from the forest to form the candidate population (*cp*) using the predefined parameters *area limit* and *lifetime*. The trees to be added to the candidate population can be done in two ways. The first step omits the trees whose age is greater than the pre-defined parameter *lifetime* indicating the normal death of trees. Even after omitting the trees that crossed the *lifetime*, if the size of the forest is greater than the predefined parameter *area limit*, the second method computes the fitness values of all the trees and is sorted in descending order. The trees having the least fitness value and are beyond the *area limit* are omitted and transferred to the candidate population, indicating the trees are died due to some genetic or environmental problems leading to the survival of the fittest in the forest. This candidate population is then used in the next phase of the global seeding of trees.

3.1.4Global seeding operation

In a GSO, a portion of the trees is selected randomly for performing a GSO. The portion of the trees is the value specified in the predefined parameter *transfer rate*. For the randomly selected portion of the trees, some of the features are randomly selected and the values of the features are flipped as in the LSO. Here the number of features to be selected for performing the GSO is the value specified in the predefined parameter *GSC*. This is carried out by changing the values of the selected features from 0 to 1 and from 1 to 0 which is similar to an LSO except that the features are processed at the same time. This implies that some features are added or some existing features are removed at the same time before applying the learning algorithm and is considered as the global search in the entire search space. The sample global seeding for the tree with four features specified in Equation 2 with GSC as 3 is shown in *Figure 3*. Since GSC is set as 3, the three features f_1, f_2 and f_4 are chosen randomly and are processed at the same time.

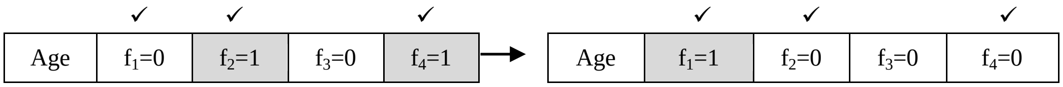


Figure 3 Simple illustration for GSO

3.1.5Update the best tree

After performing the GSO, the trees are evaluated using the fitness function and the trees are sorted based on their fitness value. The tree with the highest fitness value is considered to be the best optimal tree for which the age parameter is reset as 0 and becomes the parent for the next generation in the forest. This helps to optimize the solution effectively.

Thus, all the aforementioned stages are performed repeatedly until one of the termination criteria is fulfilled [39]. The stopping conditions include 1) a limited number of iterations 2) no changes in the evaluated fitness value for the subsequent iterations 3) a specified degree of fitness value is achieved.

3.1.6Fitness function evaluation

The proposed model uses the IRF classifier to predict the attrition results. Thus, to improve the performance of the classifier further, the performance of the classifier is taken as one of the parameters to compute the fitness function. The proposed fitness function is computed based on the classification analysis (CP_{tree}) and feature analysis (FP_{tree}) with different weights. For classification analysis, the classification accuracy, sensitivity and specificity of the classifier are considered, whereas, for the feature analysis, the mutual information between each feature with a class variable representing relevancy and the average mutual information between the feature are utilized and all the other features representing redundancy are omitted. The fitness value for the proposed model is given in Equation 6. The performance of the classification is computed by averaging the accuracy *A* that varies between 0 and 1 as well as with the product of specificity *Sp* and sensitivity *Sn* which always lies between 0 and 1.

$$fitness = W_A \times CP_{tree} + W_f \times \sum_{f_i \in tree} FP_{tree} \tag{6}$$

$$CP_{tree} = \frac{A+(Sp \times Sn)}{2} \tag{7}$$

$$FP_{tree} = NMI_{f_{i,c}} - \frac{\sum_{f_i \in tree} w_i \times NMI(f_i, f_j)}{\sum w_i} \tag{8}$$

The variable $NMI_{f_{i,c}}$ refers to the normalized mutual information (NMI) with the feature f_i and class variable *c*. This represents the level of relevancy of the feature concerning the class variable which must be always higher. The weighted average of NMI is computed between each feature with all the other features. Here, the weights refer to the IG computed at the initial phase for all the features. This step refers to the computation of redundancy which must be always minimum for the best features. Thus, the mutual information of the feature with other features is subtracted from the class variable. The variable W_A is the weight of the classification performance and it is set between 75% to 100% [40] and W_f refers to the weight of the feature which can be set as $1\text{-}W_A$ [41].

The NMI is a qualitative measure that is used to analyse the correlation of the feature with the class variable as well as with the other variables. It is computed as referred by Kachouie and Shutaywi (2020) [42]. It is often computed using entropy and conditional entropy as in Equation 9.

$$NMI(A,B) = \frac{2(H(A)-H(A|B))}{H(A)+H(B)} \quad (9)$$

In the above formula, *H(a)* and *H(b)* refer to the entropies as in Equation 3 and *H(a|b)* refers to conditional entropy and is computed as in Equation 10.

$$H(A|B) = -\sum_{a\in A, b\in B} p(a,b)\log\frac{p(a,b)}{p(a)} \quad (10)$$

The pseudocode for the proposed feature selection using an enhanced weight-based forest optimization algorithm is presented in *Figure 4*.

Algorithm: FSWFOA Algorithm

```
Input: training_set t, LSC, area_limit, life_time, transfer_rate, GSC
Output: selected optimal features having the largest fitness value
Begin
  Apply IG for all features in the t and sort them
  Form the forest with an initial tree with D features and initial elements as age
  Set the first half of the sorted features to 1 and the remaining elements to 0 for the tree
  While termination criteria are not fulfilled do
    // Local Seeding Operation
    For each zero aged trees apply LSO do
      For i ranges 1 to LSC do
        Generate a tree by selecting a random feature of the selected tree.
        Modifying feature value from 0 to 1/ 1 to 0.
        Set the age of the newly generated tree to 0.
      End For
      Increment the age of the old tree by 1.
    End For
    // Limiting the Population in the Forest
    For each tree in the forest having n trees do
      If $age > life\_time$ then
        Transfer the tree to cp and decrement n by 1.
      End If
    End For
    If $n > area\_limit$ then
      For each tree in the forest having m trees do
        Apply IRF classifier and evaluate fitness value as in Eq. (6).
      End For
      Sort the trees based on their fitness value in descending manner
      Transfer the trees beyond area_limit to cp
    End If
    // Global Seeding Operation
    Select transfer_rate percentage of trees from the cp
    For each selected tree do
      Generate a new tree similar to the selected tree
      For i ranges 1 to GSC do
        Select random features & modify their value from 0 to 1 or from 1 to 0.
      End for
    End for
    // Fitness Function Evaluation
    Apply Algorithm 2 (IRF Classifier) and evaluate its fitness value as in Eq. (6).
    Identify the tree with maximum fitness value and declare it as the best tree.
    Set the age of the best tree to 0.
  End While
  Return the best tree with features set as 1 indicating the selected optimal features
```

Figure 4 Pseudocode for proposed feature selection algorithm

3.2IRF classifier

RF is a supervised learning algorithm that builds a random number of decision trees with a random set of *n* different samples with *k* samples in the input dataset and takes the majority voting from the decision trees to classify the result. The algorithm can be used for classification as well as regression problems. Many of the existing works on the employee attrition prediction utilize various classifier algorithms among which it is proved that the RF classifier offers better results and outperforms other traditional classifiers such as KNN, LR, linear support vector machine (LSVM), NB, DT [6, 22]. However, the obtained accuracy of the RF algorithm is yet to be improved.

Thus, the proposed IRF classifier builds the random number of decision trees by selecting n number of samples from the total of k samples and the final prediction is evaluated using the performance of the classier instead of using majority voting. Several variations in predicting the class labels in the RF classifier have been proposed by researchers using weights and other techniques [43]. In the proposed model, the performance of each tree is evaluated using the classification performance, such as accuracy, specificity and sensitivity metrics as in Equation 7. Thus, the accuracy, sensitivity and specificity of each tree are computed by which the score for each tree is computed as [accuracy + (specificity × sensitivity)] /2.

Here, accuracy refers to the ability of a model to correctly predict the attrition class labels and can be computed as in Equation 11.

$$Accuracy = \frac{TruePositive+TrueNegative}{TruePositive+FalsePositive+TrueNegative+FalseNegative} \quad (11)$$

Sensitivity specifies the ability of a model to predict an employee who leaves the organization as actually left and can be computed as in Equation 12.

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \quad (12)$$

Specificity refers to the ability of the model to predict an employee who did not leave the organization as actually remain and can be computed as in Equation 13.

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \quad (13)$$

Finally, the average of the performance scores of trees for each class is computed. The class having the highest score for the performance metrics is used for classification. Thus, the class label can be predicted by grouping the decision trees based on the predicted class labels as in Equation 14.

$$CP_{tree}^{k} \in \mathbb{G}_c \; iff \; CP_{tree}^{k} \in C_i \quad (14)$$

The variable c in $\mathbb{G}_c$ represents the number of distinct classes in the dataset. Here, the computed performance analysis of the kth decision tree CP_{tree}^{k} belongs to the particular group $\mathbb{G}_c$ if it belongs to the particular class C_i. The formula to predict the class label is given in Equation 15.

$$cl(x_{test}) = cl\left(Max(\text{Avg}_c(CP_{tree}^{k}))\right) \quad (15)$$

The variable cl represents the class label and x_{test} represents the test sample x. The average performance metric specified in Equation 7 for the decision trees predicted with a specific class is computed and the class having maximum average value is the final predicted class label for the test sample x.

The main drawback of the traditional RF classifier that uses majority voting is that it treats all trees the same despite the performance of the classifier. Each model equally contributes to the prediction results even though some models poorly contribute to prediction [44, 45]. Thus, the proposed model utilizes evaluation metrics such as accuracy, specificity and sensitivity for assigning scores or weights to the models. These values are then averaged based on the predicted class labels of the trees and finally averaged in predicting the final results.

The algorithm pseudocode for the IRF Classifier is presented in *Figure 5*.

```
Algorithm: IRF Classifier
Input: training set D with M features, number of trees T, test instance x
Output: Predicted class for the given instance x
Procedure IRF_Classifier
Begin
  For T from 1 to t do
      Randomly select the training set D' from D
      Randomly select the M' from M features and store them in feature_list
      Call Procedure Generate_trees(M', D', x)
  End for
  For T from 1 to t do
      Predict the class label for the test instance x
```

```
    Evaluate accuracy, sensitivity and specificity
    Compute performance $CP_{tree}$ as in Eq. (7)
End For
Calculate the score for each class c by averaging $CP_{tree}$ for tree t resulting in class c
Identify $c_i$ having a max. the score for the performance
Return $c_i$ as the predicted label for the instance x.

Procedure Generate_trees(M', D'):
    Create node N containing D'
    If instances of D' are from the same class then
        return N as a leaf node labelled with the class C
    End If
    Apply information gin for the features
    Select the feature F with the highest IG as a split node
    Create k child nodes $N_1$, $N_2$, …, $N_f$ where F has f possible values $F_1$, $F_2$, .., $F_f$
    feature_list = feature_list-split_ feature
    For each child node k do
        $D_i$ be the set of instances in D' satisfying split criterion of k
        If $D_i$ is empty then
            Attach a leaf node to N labelled with class C having max. instances
        Else
            Leaf node is Generate_tree($D_i$, attribute list) to node N;
        End If
    End For
Return N
```

Figure 5 Algorithm pseudocode for proposed IRF classifier

4.Results

This section presents the details of the experimental analysis performed for the proposed model along with the dataset used and the result analysis as well as the comparison of results with existing classifiers. An experimental setup has been made for implementing the model which is carried out on the system with the following configuration, such as Intel(R) Core (TM) i3-4005U CPU with 1.70 GHz, 4 GB RAM and 64-bit Windows Operating System. The software used for writing the code is Python 3.5 in Anaconda Environment with Jupyter Notebook. For analysing the features, free tools such as WEKA and ORANGE are used for performing machine learning tasks as well as data analysis and visualisation respectively.

4.1Dataset used

For predicting employee attrition using the proposed weighted forest optimization algorithm with an IRF classifier, the IBM HR analytics employee attrition dataset from Kaggle [46]. The dataset contains instances that represent the details about the employees for the HR analytics case study. The target variable is attrition in which the value 0 represents the employee who did not leave and the value 1 represents the employee who left the company. The dataset contains 35 features including class variable and 1470 instances with 237 instances with the target variable as 'yes' and 1233 instances with the target variable as 'no'.

4.2Model preparation

Model preparation is an initial stage in performing the analysis of the input data. Here the input dataset undergoes various pre-processing steps to make the data suitable for applying the proposed techniques and to get a better result [47]. In general, several steps exist in data mining such as data cleaning, data transformation and data reduction [38]. Data cleaning deals with the missing and incomplete values in the dataset by either filling the missing data with appropriate values or deleting the records. However, for the analysis of the proposed model, as the dataset contains no missing values and therefore this step is skipped [48]. The next step is data transformation which deals with structuring the values that is suitable for processing easily.

Data discretization is one such method that converts the given set of large attribute values to smaller sets without losing data. More specifically, the huge set of continuous numeric data is converted to a finite set of intervals. In the proposed model, discretization is carried out using an instance, filtering technique called binning which converts the huge set of numeric data into nominal values [49]. Data normalization is another transformation technique that intends to scale down the huge range of values into a smaller range which is highly significant since a different range of values for different features often produce poor results. In the proposed model, min-max normalization is used that linearly transforms the existing range to a new range based on the minimum and maximum values in the range [50]. The formula for min-max normalization is given in Equation 16.

$$\mathrm{Val}_{\mathrm{scaled}} = \frac{\mathrm{Val} - \mathrm{Mn}_{\mathrm{A}}}{\mathrm{Mx}_{\mathrm{A}} - \mathrm{Mn}_{\mathrm{A}}}(\mathrm{NewMx}_{\mathrm{A}} - \mathrm{NewMn}_{\mathrm{A}}) + \mathrm{NewMn}_{\mathrm{A}} \quad (16)$$

The next step is data reduction where it intends to reduce the data to improve the prediction results. In the proposed model, the enhanced weighted forest optimization algorithm is applied to select the significant features that contribute more to the classification performance. The proposed model selects 16 vital features to predict attrition results.

4.3Evaluation metrics

For any proposed model to be proved efficient, it has to be evaluated using a series of performance metrics [51]. The proposed model utilizes various metrics such as accuracy, error rate, true positive (TP) rate, false positive (FP) rate, precision, sensitivity, and f-measure for performance analysis. In common, accuracy is the most important metric used in evaluating the prediction models [52]. It is the ratio of the number of correctly predicted results with the total number of predictions made. The error rate is the ratio of the number of incorrectly predicted results to the total number of predictions carried out. Precision is the ratio of correctly predicted positive cases with the total positive predictions, whereas, sensitivity or recall is the ratio of correctly predicted positive cases with the total actual positive cases. F-measure evaluates the model by applying geometric mean to the precision and recall values.

4.4Exploratory data analysis

For the attrition dataset used in the proposed study, an extensive exploratory data analysis has been made. Accordingly, the dataset contains 83.88% of records with attrition results as 'No' and 16.12% of instances with attrition values as 'Yes'. Also, a detailed analysis of the features has been made for which the features are categorized into three basic types 1) Demographic (D), 2) Behavioural (B) and 3) Attitudinal (A) [7]. Demographic features represent a set of socioeconomic information of an employee, including age, gender, education, income and more. Behavioural features describe the observed actions of an employee and attitudinal data refers to the opinions and preferences of a person. Also, the attribute values of the features are categorized as numeric and categorical. The details about the list of features, attribute value description for each feature, the categorization and selected features are presented in *Table 1*.

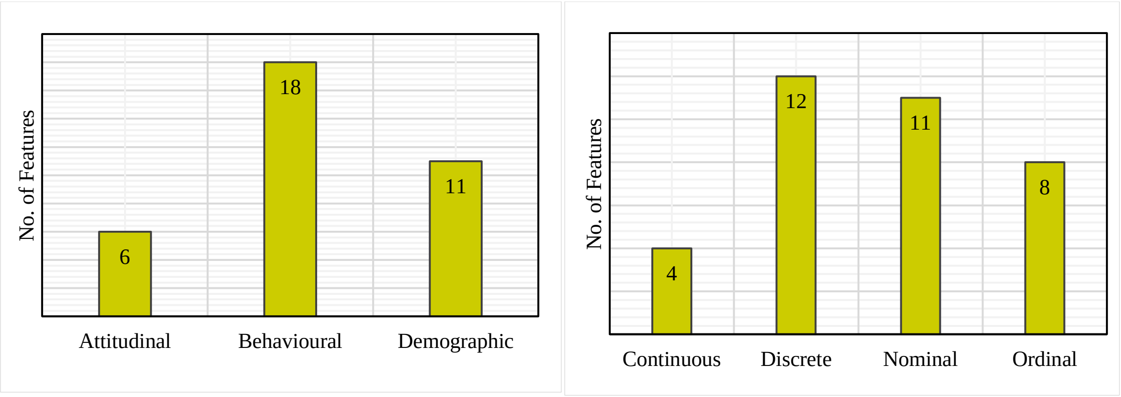
A detailed analysis of the categorization of the features is made and the results are represented as a graph in *Figure 6*. Consequently, 51.4% of the features belong to the behavioural category which is more dominant, whereas demographic and attitudinal features correspond to 31.4% and 17.14% respectively. Similarly, with variable categorization, most of the features are discrete and nominal than ordinal and continuous.

The datatype categorization is also made for the features that include more integer types than string and Boolean values. By using the proposed enhanced weighted forest optimization for selecting significant features, 16 features are selected for further classification using an IRF classifier. The list of selected features based on their significance level is Over_Time, Job_Level, Monthly_Income, Total_Working_Years, Years_At_Company, Years_With_Curr_Manager, Age, Stock_Option_Level, Martial_Status, Job_Satisfaction, Environment_Satisfaction, Job_Role, Business_Travel, Job_Involvement, Work_Life_Balance and Gender. Thus, the factors such as working overtime, the level of Job, and monthly income are the major cause for the employee to leave the organization. While analysing the feature categorization, about 83% of selected features are attitudinal, 45% of selected features are demographic and the remaining 33% of selected features is behavioural based. The analysis of feature categorization on selected features is represented as a graph in *Figure 7*.

Table 1 List of features in attrition dataset and its categorization

S. No.	List of features	Attribute value description	Feature type	Data type	Selected feature
1	Age	18-60	Demographic	Numeric	✓
2	Business_Travel	No Travel, Travel Frequently, Travel Rarely	Behavioural	Categorical	✓
3	Daily_Rate	Salary Level ranges from 102 to 1499	Behavioural	Numeric	
4	Department	Human Resource, Research and Development, Sales	Demographic	Categorical	

S. No.	List of features	Attribute value description	Feature type	Data type	Selected feature
5	Distance_From_Home	The distance from home to work ranges from 1 to 29	Demographic	Numeric	
6	Education	1=Below College, 2=College, 3=Bachelor, 4=Master, 5=Doctor	Demographic	Categorical	
7	Education_Field	1=Human Resources, 2=Life Science, 3=Marketing, 4=Medical, 5= Others, 6=Technical Degree	Demographic	Categorical	
8	Employee_Count	1	Behavioural	Numeric	
9	Employee_Number	Employee ID ranges from 1 to 2068	Demographic	Numeric	
10	Environment_Satisfaction	1=Low, 2=Medium, 3=High, 4=Very High	Attitudinal	Categorical	✓
11	Gender	Male, Female	Demographic	Categorical	✓
12	Hourly_Rate	Hourly Salary ranges from 30 to 100	Behavioural	Numeric	
13	Job_Involvement	1=Low, 2=Medium, 3=High, 4=Very High	Attitudinal	Categorical	✓
14	Job_Level	The level of Job ranges from 1 to 5	Attitudinal	Categorical	✓
15	Job_Role	Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Human Resources, Sales Representative, Manager, Research Director	Demographic	Categorical	✓
16	Job_Satisfaction	1=Low, 2=Medium, 3=High, 4=Very High	Attitudinal	Categorical	✓
17	Marital_Status	Single, Married, Divorced	Demographic	Categorical	✓
18	Monthly_Income	The monthly Salary ranges from 1009 to 20000	Demographic	Numeric	✓
19	Monthly_Rate	The monthly rate ranges from 2094 to 27000	Behavioural	Numeric	
20	Num_Companies_Worked	The number of companies, worked at ranges from 0 to 9	Behavioural	Numeric	
21	Over18	Yes, No	Demographic	Categorical	
22	Over_Time	Yes, No	Behavioural	Categorical	✓
23	Percent_Salary_Hike	Percentage increase in salary ranges from 11 to 25	Behavioural	Numeric	
24	Performance_Rating	1=Low, 2=Good, 3=Excellent, 4=Outstanding	Behavioural	Categorical	
25	Relationship_Satisfaction	1=Low, 2=Medium, 3=High, 4=Very High	Attitudinal	Categorical	
26	Standard_Hours	Standard working hours are 80	Behavioural	Numeric	
27	Stock_Option_Level	Stock options range from 0 to 3	Behavioural	Numeric	✓
28	Total_Working_Years	Total years worked range from 0 to 40	Behavioural	Numeric	✓
29	Training_Times_Last_Year	Hours spent in training range from 0 to 6	Behavioural	Numeric	
30	Work_Life_Balance	1=Bad, 2=Good, 3=Better, 4=Best	Attitudinal	Categorical	✓
31	Years_At_Company	Total years worked at the company range from 0 to 40	Behavioural	Numeric	✓
32	Years_In_Current_Role	Years in current role range from 0 to 18	Behavioural	Numeric	
33	Years_Since_Last_Promotion	Last promotion ranges from 0 to 15	Behavioural	Numeric	
34	Years_With_Current_Manager	Years spent with a current manager range from 0 to 17	Behavioural	Numeric	✓
35	Attrition (Target)	Yes, No	Behavioural	Categorical	✓



a) Feature categorization b) Feature categorization

Figure 6 Feature analysis of employee attrition dataset

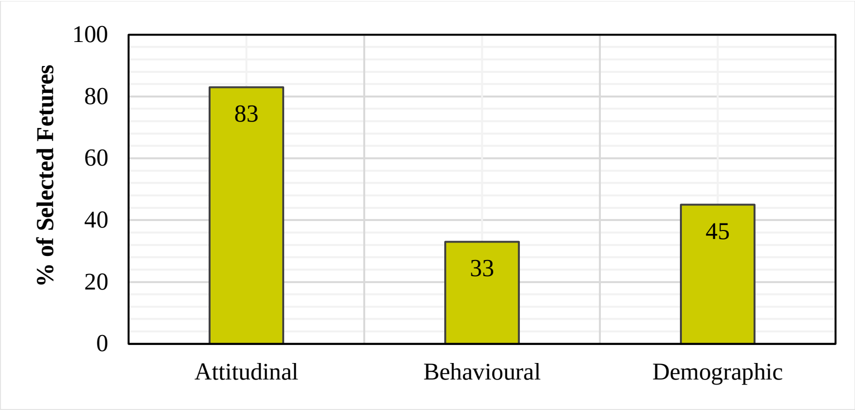
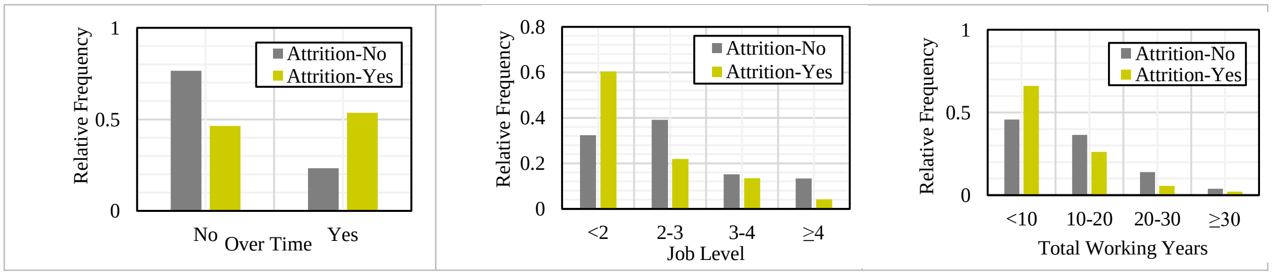


Figure 7 Categorization analysis of selected feature

Thus, from the analysis, it is clear that the attitudes of the employees and employers play a crucial role in employee Attrition. Apart from attitudinal features, demographic information of the employee also significantly affects the attrition decision of the employee than behavioural based information. Furthermore, a more detailed assessment has been made on the attribute values of the selected features to identify the significant reason for employee attrition. The detailed analysis of the variables for the selected features is presented as a graph in *Figure 8*. In the graph, the x-axis represents variable values of the selected features and the y-axis represents the relative frequency of the values in the dataset with respect to the class variable.

From the analysis, various possibilities exist that make the employee leave an organization have been uncovered. The foremost reason for employee attrition is overtime. The employee doing overtime has a 90% possibility to quit the job. The level of job is also playing a significant role in employee attrition in which the low-level positions in the job hierarchy have the possibility of 74% resigning their jobs. The employee gets a monthly salary of less than 4K and having minimum total work experience, experience at the company less than 10 years as well as minimum working years with the current manager has the highest possibility of 78%, 78% 82% and 81% respectively to quit their job. The attrition rate will be more common among the workers with an age <23, between 23-28 and 28-34 which has the possibility of 56%, 77% and 78% respectively.

The workers with minimum stock level, workers with marital status single and mere satisfaction in their job and working environment has the highest possibility of 76%, 75%, 77% and 75% to switch their job repeatedly. Among various job roles considered in the dataset including healthcare representative (HIR), HR, manager (Man), manufacturing director (MD), research director (RD), research scientist (RS), laboratory technician (LT), sales executive (SE) and sales representative (SR), the employees at low job positions such as LT, SE, SR has the highest possibility of 76%, 82% and 77% to leave the company. The workers who encounter frequent business travel, less job involvement, minimum work-life balance and gender as male has more chance to leave their job with the possibility of 75%, 78%, 80% and 83% respectively.



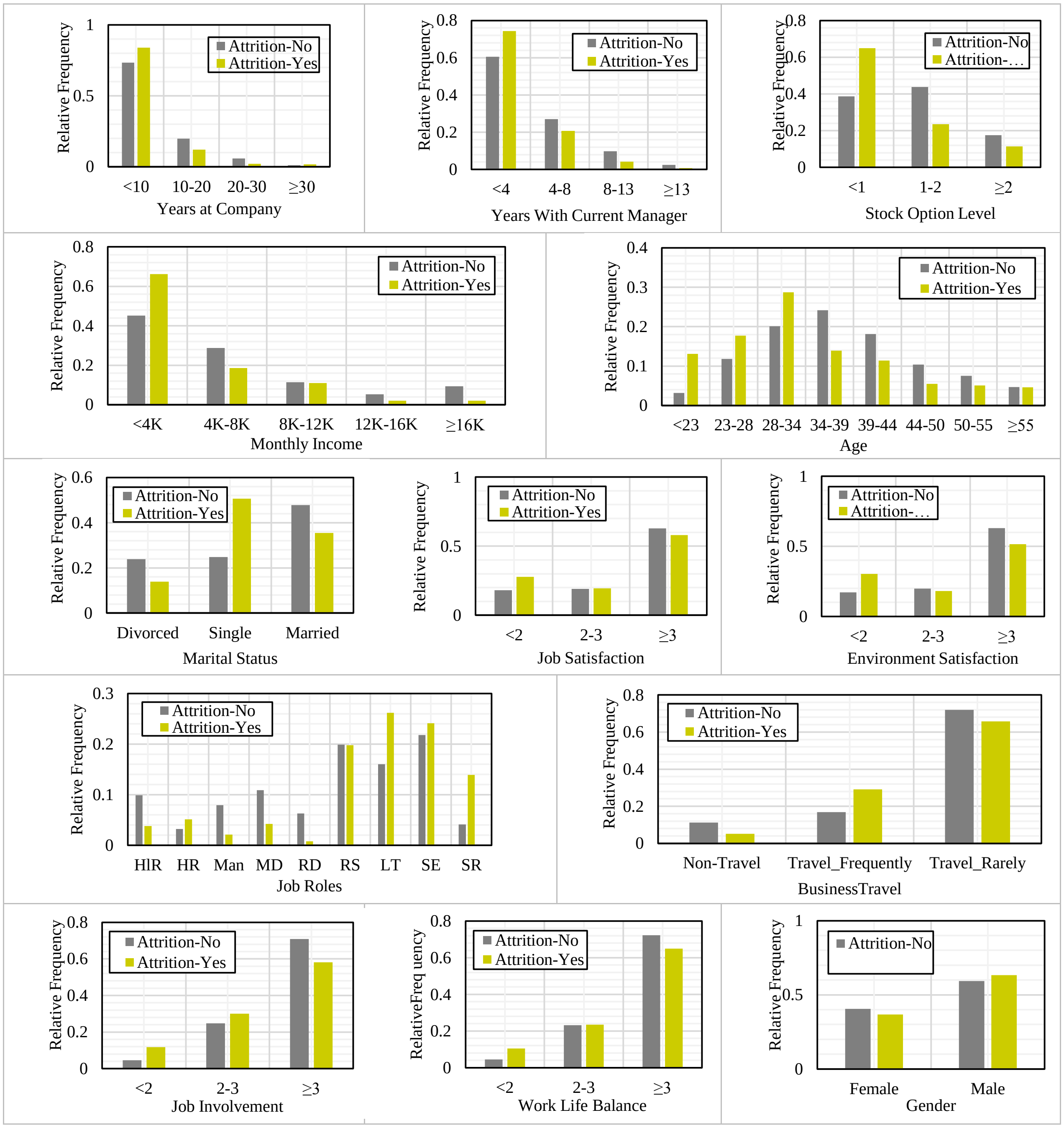


Figure 8 Variable analysis of the selected features

4.5 Performance analysis

Performance analysis has been made for the proposed IRF classifier as well as an enhanced weighted forest optimization algorithm used for feature selection. Initially, the performance of the proposed IRF classifier is compared with various other standard classifiers. Some of the existing classifiers used for the analysis are NB, LR, SVM, and KNN, AdaBoost (AB), bagging (Bag), C4.5 tree (C4.5), logistic model tree (LMT), and RF were executed with 10-fold cross-validation. Various performance metrics such as accuracy (Acc.), error rate (ER), precision (Pre.), sensitivity (Sen.), and F-measure (FM) for the proposed and existing classification models are compared and the results are presented in *Table 2*. The obtained results are then used for analysing the performance of the proposed classifier and so, all 35 features are used for analysis.

The obtained results with the metrics such as accuracy, precision, sensitivity and f-measure for the existing and proposed models in the analysis are presented as a graph for understanding in *Figure 9.*

Table 2 Performance comparison of proposed IRF classifier

Classifier	Acc. (%)	ER (%)	Pre. (%)	Sen. (%)	FM (%)
NB	81.50	18.50	82.90	81.50	82.10
LR	85.44	14.56	84.70	85.40	85.00
SVM	85.71	14.29	84.80	85.70	**85.10**
KNN	78.38	21.63	77.80	78.40	78.10
AB	85.85	14.14	85.00	85.90	81.70
Bag	83.74	16.26	83.70	83.70	77.50
C4.5	84.01	15.99	80.60	84.00	81.00
LMT	86.94	13.06	85.40	86.90	84.70
RF	84.49	15.51	85.50	84.50	78.10
IRF	**88.85**	**11.15**	**88.80**	**88.90**	84.40

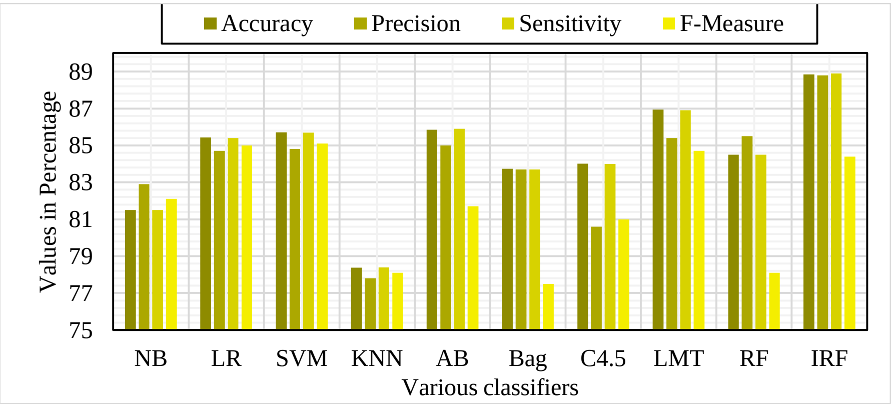


Figure 9 Performance comparison of different classifiers

Upon analysing the obtained results, it is clear that the proposed IRF model has better accuracy of about 88.85% and a minimum error rate of 11.15% than other existing models under comparison. Similarly, the TP rate, precision and sensitivity of the proposed classifier are also showing improved results of 88.9%, 88.8% and 88.9% respectively. However, for the FP rate, the NB classifier has a minimum value of 42.40% than the proposed classifier which has a value of 72.60%. Also, the f-measure for the proposed model has 84.40%, which is a little lower than the SVM classifier. Thus, it can be concluded that the proposed IRF classifier offers improved performance of about 71.48% than other models under comparison.

The complete proposed model FSWFOA has been evaluated using various performance metrics such as accuracy (Acc.), error rate (ER), precision (Pre.) and sensitivity (Sen.) and the obtained results are compared with various feature selection models including wrapper approach using RF, KNN and other rank based filtering approaches such as correlation based feature subset selection (CFS), IG, principle component analysis (PCA), relief (Rel.) algorithm, symmetrical uncertainty (SU), gain ratio (GR) and FSFOA. Each of the feature selection techniques is evaluated using various classifiers such as SVM, LR and RF. For identifying interesting features using RF, an ensemble model is also used for the analysis [9]. The proposed enhanced weighted forest optimization algorithm is also implemented with the proposed IRF classifiers and other classifiers. The attained outcomes are displayed in *Table 3*. The table presents the feature selection technique used along with the number of features selected by the technique and the corresponding classifiers used in the analysis.

Table 3 Performance comparison of proposed model

Feature selection model	#Feature	Classifier	Acc.	ER	Pre.	Sen.
RF	10	SVM	77.65	22.35	84.95	86.16
		LR	81.77	18.23	93.39	83.21
		RF	82.64	17.36	88.93	88.14
		Ensemble	83.87	16.13	93.17	87.17
KNN	14	SVM	83.88	16.12	83.9	83.9
		LR	85.58	14.42	83.4	85.6
		RF	85.65	14.35	83.4	85.6
CFS	9	SVM	86.12	13.88	84.7	86.1
		LR	86.67	13.33	85.0	86.7
		RF	84.08	15.92	80.7	84.1
IG	12	SVM	83.87	16.13	83.9	83.9
		LR	85.57	14.43	83.3	85.6
		RF	85.37	14.63	83.0	85.4
PCA	15	SVM	83.88	16.12	83.9	83.9
		LR	84.69	15.31	84.0	84.7
		RF	84.56	15.44	81.3	84.6
Rel.	13	SVM	83.94	16.06	86.5	83.9
		LR	86.46	13.54	84.7	86.5
		RF	85.17	14.83	82.7	85.2
SU	14	SVM	83.88	16.12	83.9	83.9
		LR	86.26	13.74	84.4	86.3
		RF	86.25	13.75	84.4	86.3
GR	11	SVM	83.88	16.12	83.9	83.9
		LR	86.26	13.74	84.4	86.3
		RF	85.64	14.36	83.4	85.6
FSFOA	17	SVM	87.88	12.12	87.9	87.9
		LR	89.23	10.77	86.8	89.2
		RF	88.83	11.17	86.5	88.8
FSWFOA (Proposed)	15	SVM	87.8	12.2	87.9	88.7
		LR	89.92	10.08	87.9	89.9
		RF	89.44	10.56	87.0	89.4
		IRF (Proposed)	91.23	8.77	89.52	91.17

Specifically, with FOA for feature selection, the model provides good results with the LR model of about 89.23% of accuracy. However, the proposed enhanced weighted forest optimization algorithm offers a higher performance rate with LR and IRF classifiers of about 89.92% and 91.23% respectively. The results obtained from the optimization models such as FSFOA and FSWFOA with various classifiers such as SVM, LR, RF and IRF are presented as a graph in *Figure 10.*

Also, to analyse the performance of the model concerning the resources used, the time consumption and RAM utilized by the proposed model are evaluated and the results are compared with other models [27] such as traditional C5.0, and C5.0 with association rule, PSO optimized, GWO optimized. The performance comparison on resource utilization is shown in *Figure 11.*

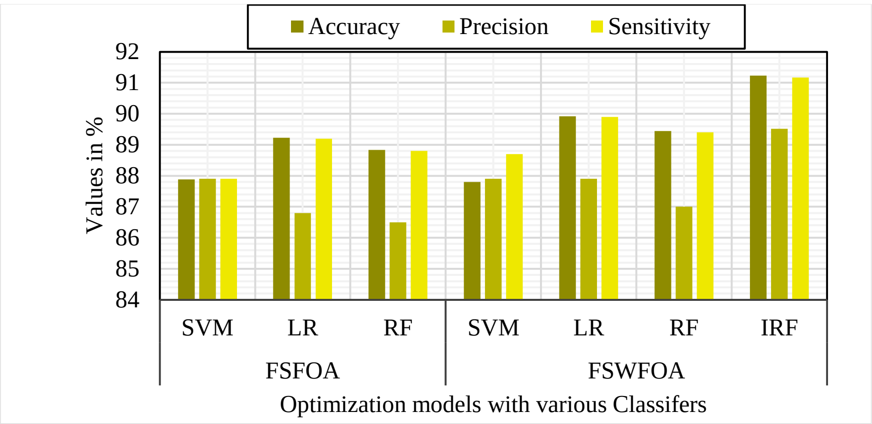
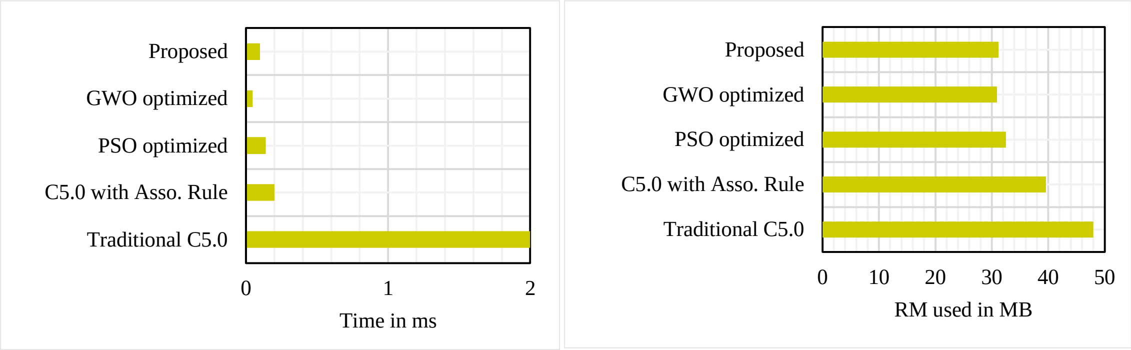


Figure 10 Performance comparison of optimization models



a) Time consumed b) RAM consumed

Figure 11 Performance comparison of resources utilized

The proposed model consumes 0.098 ms of time to complete the process and 31.2MB of RAM, which is lower than most of the other models except GWO optimized which takes 0.046 ms time and 30.9MB RAM of memory.

It is clear that the proposed model results in higher accuracy, precision and sensitivity, which are highly necessary for any classifier model in predicting accurate values with minimum resource utilization of memory and time.

5.Discussion

From the analysis of the dataset, it is clear that the behaviour related and demographic features are utilized more than attitudinal features in predicting employee attrition. However, upon applying the feature selection algorithm, it is found that the factors such as working overtime, the level of Job, and monthly income are the primary reason for the employee to leave the company. Moreover, while evaluating the feature categorization, it is surprised to see that attitudinal features contribute more than demographic and behavioural based features. Thus, it is identified that the attitude of the employee plays a more primary role in employee attrition than the socioeconomic characteristics and behaviours of an employee.

On analyzing the IRF classifiers with the conventional classifiers, it is clear that the proposed IRF model has better accuracy with a minimum error rate. Similarly, the TP rate, precision and sensitivity of the proposed classifier are also showing improved results. Though the proposed model offers lower performance with respect to FP rate and F-measure, it is still offering better values for the other 5 metrics out of 7 metrics used in the analysis. Thus, it can be concluded that the proposed IRF classifier offers improved performance than other models under comparison.

The analysis of various feature selection techniques shows that the correlation-based feature subset selection offers better results for SVM and LR classifiers. Gain ratio and Relief based feature selection offers good results with LR whereas symmetrical uncertainty-based feature selection provides better results with RF and LR classifiers. In comparison with the machine learning techniques for feature selection, optimization algorithms offer better results. The proposed optimization based feature selection FSWFOA with various classifiers such as SVM, LR, RF and IRF has been analysed in which the proposed IRF classifier offers higher performance than other models used in the analysis. It is also proved that the proposed model also offers better results concerning the utilization of resources such as RAM and time consumption.

With detailed analysis, some of the research implications have been drawn in which the foremost reason for employee attrition is overtime. Thus, there exists a high correlation between the overtime and the attrition rate. The low carter employees having low-level positions in the job hierarchy, such as Laboratory Technicians, Sales Executive and Sales Representative with salary less than 4K has a high probability of leaving the company. Thus, there exists a high-level dependency between the low positions jobs with minimum salary and the attrition rate. Similarly, the employees with minimum total work experience, experience at the company less than 10 years as well as minimum working years with the current manager have the highest possibility to quit their job. Moreover, the attrition rate is high for the employees with an age less than 34 with marital status single and minimum satisfaction in their job and working environment has the highest possibility to switch their job repeatedly. The male employees encountering frequent business travel, less job involvement, and minimum work-life balance is more likely to leave their job.

Based on the knowledge obtained from the empirical analysis, strategic planning can be made to retain the talent. Talent management is a significant field for the improvement of the organization. Some of the points that are suggested to be included in the strategic plan are:

- Employing the precise talent for the exact position.
- Encouraging to work only in working hours and adequate rest hours.
- Providing opportunity and necessary resources to grow along with the organization.
- Recognizing the talents and offering flexibility, promotions, awards and incentives regularly.
- Providing a better employee stock option plan to make the employee an equity holder.
- Encouraging healthy work-life balance.
- Offering a competitive base salary and other benefits.
- Maintaining a healthy relationship with the employees.
- Offering business travel only based on their willingness.
- Creating a safe environment and gaining trust.

5.1Limitation of the study

Like any research, the proposed research study has some limitations. First, the features used for the analysis are confined to the dataset used in the study created by IBM. However, apart from these features, several other features such as workload, chances for a job transfer, and transport may also influence the employee to leave the current organization. The study suggests for identification of all possible core factors to be considered in predicting employee attrition. Second, the study focuses mainly on proposing the automated system in attrition prediction and does not consider the managerial aspects. A detailed study of management-based approaches is necessary to manage and retain the talents by constituting appropriate strategic plans and policies for employees, management and organization in the real-time environment. A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future work

This paper presents the two phase prediction model to forecast employee attrition in which the first phase delivers an enhanced weighted forest optimization algorithm for selecting the substantial features to improve the classification accuracy and the second phase with a modified RF classifier for managing talents. The model has been evaluated using IBM human resource employee attrition dataset with 35 features in which the proposed model identifies 16 significant features. By performing detailed exploratory analysis, it is identified that the attitudinal based features contribute more in employee attrition than demographic and behavioural based features. Also, the identified significant features are evaluated in which a few reasons for attrition have been identified for which some of the strategic planning are also suggested. The performance analysis has been made and the results are evaluated in which the IRF model offers 88.85% of accuracy and 88.90% of TP rate. The FSWFOA with IRF model provides improved results with an accuracy of 91.23% and a minimum error rate of 8.77% than several other models used for the comparison. The future work focuses on improving the accuracy of the model to 100% and ascertaining better strategic plans to retain the talents in the real-time environment.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contributions statements

S. Porkodi: Conceptualization and design of the work, data collection, writing and editing original draft. **S. Srihari:** Conceptualization and design of the work, result analysis, implementation, interpretation and review of results. **N. Vijayakumar:** Conceptualization and design of the work, implementation, investigation on the result analysis. All authors discussed the results, contributed to the final manuscript and provided critical feedback and helped shape the research, analysis and manuscript.

References

[1] Schweyer A. Talent management systems: best practices in technology solutions for recruitment, retention and workforce planning. John Wiley & Sons; 2004.

[2] Castellano WG. Practices for engaging the 21st century workforce: challenges of talent management in a changing workplace. FT Press; 2013.

[3] Lewis RE, Heckman RJ. Talent management: a critical review. Human Resource Management Review. 2006; 16(2):139-54.

[4] Dibble S. Keeping your valuable employees: retention strategies for your organization's most important resource. John Wiley & Sons; 1999.

[5] Abdurakhmanova G, Shayusupova N, Irmatova A, Rustamov D. The role of the digital economy in the development of the human capital market. The Research Archive. 2020; 24(7): 8043-51.

[6] Sisodia DS, Vishwakarma S, Pujahari A. Evaluation of machine learning models for employee churn prediction. In international conference on inventive computing and informatics 2017 (pp. 1016-20). IEEE.

[7] Maisuradze M. Predictive analysis on the example of employee turnover. Tallinn University of Technology. 2017.

[8] Nocker M, Sena V. Big data and human resources management: the rise of talent analytics. Social Sciences. 2019; 8(10):1-19.

[9] Karande S, Shyamala L. Prediction of employee turnover using ensemble learning. In ambient communications and computer systems 2019 (pp. 319-27). Springer, Singapore.

[10] Yadav S, Jain A, Singh D. Early prediction of employee attrition using data mining techniques. In international advance computing conference 2018 (pp. 349-54). IEEE.

[11] Poornappriya TS, Gopinath R. Employee attrition in human resource using machine learning techniques. Webology. 2021; 18(6):2844-56.

[12] Bama SS, Saravanan A. Efficient classification using average weighted pattern score with attribute rank based feature selection. International Journal of Intelligent Systems and Applications. 2019; 11(7):29-42.

[13] Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: a review. Journal of King Saud University-Computer and Information Sciences. 2022; 34(4):1060-73.

[14] Uthayakumar J, Metawa N, Shankar K, Lakshmanaprabu SK. Financial crisis prediction model using ant colony optimization. International Journal of Information Management. 2020; 50:538-56.

[15] Xiao L. Optimal allocation model of enterprise human resources based on particle swarm optimization. In international conference on computer information and big data applications 2020(pp. 249-53). IEEE.

[16] Sankhwar S, Gupta D, Ramya KC, Sheeba RS, Shankar K, Lakshmanaprabu SK. Improved grey wolf optimization-based feature subset selection with fuzzy neural classifier for financial crisis prediction. Soft Computing. 2020; 24(1):101-10.

[17] Moorthy U, Gandhi UD. Forest optimization algorithm-based feature selection using classifier ensemble. Computational Intelligence. 2020; 36(4):1445-62.

[18] Razmjooy N, Sheykhahmad FR, Ghadimi N. A hybrid neural network–world cup optimization algorithm for melanoma detection. Open Medicine. 2018; 13(1):9-16.

[19] Pereira GT, Santos BZ, Cerri R. A genetic algorithm for transposable elements hierarchical classification rule induction. In congress on evolutionary computation 2018 (pp. 1-8). IEEE.

[20] Chiclana F, Kumar R, Mittal M, Khari M, Chatterjee JM, Baik SW. ARM–AMO: an efficient association rule mining algorithm based on animal migration optimization. Knowledge-Based Systems. 2018; 154:68-80.

[21] Kamath DR, Jamsandekar DS, Naik DP. Machine learning approach for employee attrition analysis. International Journal of Trend in Scientific Research and Development. 2019:62-7.

[22] Zhao Y, Hryniewicki MK, Cheng F, Fu B, Zhu X. Employee turnover prediction with machine learning: a reliable approach. In proceedings of SAI intelligent systems conference 2018 (pp. 737-58). Springer, Cham.

[23] Jain R, Nayyar A. Predicting employee attrition using xgboost machine learning approach. In international conference on system modeling & advancement in research trends 2018 (pp. 113-20). IEEE.

[24] Pratt M, Boudhane M, Cakula S. Employee attrition estimation using random forest algorithm. Baltic Journal of Modern Computing. 2021; 9(1):49-66.

[25] Salunkhe TP. Improving employee retention by predicting employee attrition using machine learning techniques (Doctoral dissertation, Dublin Business School). 2018.

[26] Bindra H, Sehgal K, Jain R. Optimisation of C5. 0 using association rules and prediction of employee attrition. In international conference on innovative

computing and communications 2019 (pp. 21-9). Springer, Singapore.

[27] Sehgal K, Bindra H, Batra A, Jain R. Prediction of employee attrition using GWO and PSO optimised models of C5. 0 used with association rules and analysis of optimisers. In innovations in computer science and engineering 2019 (pp. 1-8). Springer, Singapore.

[28] Jain PK, Jain M, Pamula R. Explaining and predicting employees' attrition: a machine learning approach. SN Applied Sciences. 2020; 2(4):1-11.

[29] Srivastava PR, Eachempati P. Intelligent employee retention system for attrition rate analysis and churn prediction: an ensemble machine learning and multi-criteria decision-making approach. Journal of Global Information Management. 2021; 29(6):1-29.

[30] Devi GD, Kamalakkannan S. Prediction of job satisfaction from the employee using ensemble method. In international conference on advanced computing technologies and applications 2022 (pp. 1-8). IEEE.

[31] Sharma MK, Singh D, Tyagi M, Saini A, Dhiman N, Garg R. Employee retention and attrition analysis: a novel approach on attrition prediction using fuzzy inference and ensemble machine learning. Webology. 2022; 19(2):5338-58.

[32] Rombaut E, Guerry MA. The effectiveness of employee retention through an uplift modeling approach. International Journal of Manpower. 2020; 41(8):1199-220.

[33] Wijaya D, DS JH, Barus S, Pasaribu B, Sirbu LI, Dharma A. Uplift modeling VS conventional predictive model: areliable machine learning model to solve employee turnover. International Journal of Artificial Intelligence Research. 2021; 5(1):53-64.

[34] Nagadevara V. Prediction of employee attrition using work-place related variables. 2012.

[35] Singh M, Varshney KR, Wang J, Mojsilovic A, Gill AR, Faur PI, et al. An analytics approach for proactively combating voluntary attrition of employees. In international conference on data mining workshops 2012 (pp. 317-23). IEEE.

[36] Ghaemi M, Feizi-derakhshi MR. Forest optimization algorithm. Expert Systems with Applications. 2014; 41(15):6676-87.

[37] Ghaemi M, Feizi-derakhshi MR. Feature selection using forest optimization algorithm. Pattern Recognition. 2016; 60:121-9.

[38] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

[39] Mohanty F, Rup S, Dash B, Majhi B, Swamy MN. Mammogram classification using contourlet features with forest optimization-based feature selection approach. Multimedia Tools and Applications. 2019; 78(10):12805-34.

[40] Huang CL, Wang CJ. A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications. 2006; 31(2):231-40.

[41] Zhao M, Fu C, Ji L, Tang K, Zhou M. Feature selection and parameter optimization for support vector machines: a new approach based on genetic algorithm with feature chromosomes. Expert Systems with Applications. 2011; 38(5):5197-204.

[42] Kachouie NN, Shutaywi M. Weighted mutual information for aggregated kernel clustering. Entropy. 2020; 22(3):1-15.

[43] Selvam RRLP, Saleem IAM, Alenezi A. Classification of imbalanced class distribution using random forest with multiple weight based majority voting for credit scoring. International Journal of Recent Technology and Engineering. 2019; 7(6S5):517-26.

[44] Zhang C, Wang X, Chen S, Li H, Wu X, Zhang X. A modified random forest based on kappa measure and binary artificial bee colony algorithm. IEEE Access. 2021; 9:117679-90.

[45] Birant KU. Multi-view rank-based random forest: a new algorithm for prediction in eSports. Expert Systems. 2022; 39(2).

[46] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset. Accessed 26 November 2021.

[47] Bama SS, Ahmed MI, Saravanan A. A mathematical approach for mining web content outliers using term frequency ranking. Indian Journal of Science and Technology. 2015; 8(14): 1-5.

[48] García S, Luengo J, Herrera F. Data preprocessing in data mining. Cham, Switzerland: Springer International Publishing; 2015.

[49] Gopalsamy A, Radha B. Feature selection using multiple ranks with majority vote-based relative aggregate scoring model for Parkinson dataset. In proceedings of international conference on data science and applications 2022 (pp. 1-19). Springer, Singapore.

[50] Jo JM. Effectiveness of normalization pre-processing of big data to the machine learning performance. The Journal of the Korea Institute of Electronic Communication Sciences. 2019; 14(3):547-52.

[51] Bama SS, Ahmed MI, Saravanan A. A survey on performance evaluation measures for information retrieval system. International Research Journal of Engineering and Technology. 2015; 2(2):1015-20.

[52] Sathya BS, Ahmed I, Saravanan A. Relevance re-ranking through proximity based term frequency model. In international conference on ICT innovations 2016 (pp. 219-29). Springer, Cham.

Dr. S. Porkodi is currently employed as faculty at the University of Technology and Applied Sciences (HCT), Muscat, Sultanate of Oman. She holds academic degrees in M.B.A., M.Phil(Ent)., M.Phil(Mgt)., PG.D. PM&IR., PG.D.HM., PG.D.EM., Ph.D. She has authored 11 books in the fields of Hospital administration and Business studies. She has contributed to prestigious periodicals, and published numerous papers in reputed international and national

journals. She served as an Editor of a refereed Journal-Management Stream, Editor-in-Chief of International Journal of Management Rivulet, and Editor of International Journal of Management, Entrepreneurship and Technology, Newyork.
Email: dr.porkodi@gmail.com

S. Srihari completed his Bachelor of Technology in Computer Science and Engineering at Amrita Vishwa Vidhyapeetham, Coimbatore in 2021 and is currently pursuing his Master's in Data Science at Illinois Institute of Technology, Chicago. He is deeply interested in Machine Learning, Data Science and effects of Quantum Machine Learning in the aspects of Drug Discovery.
Email: srhr1999@gmail.com

Dr. N. Vijayakumar completed Ph.D. form Bharathiyar University, Coimbatore in the field of security in cloud computing in 2020. Presently he is working as IT Instructor in Technical Administrative Training Institute Muscat Sultanate of Oman.

Email: dr.vijayakumarnatarajan@gmail.com

Appendix I

S. No.	Abbreviation	Description
1	AB	AdaBoost
2	ACO	Ant Colony Optimization
3	AMO	Animal Migration Optimization
4	CFS	Correlation Based Feature Subset Selection
5	DT	Decision Tree
6	ER	Error Rate
7	FM	F-Measure
8	FOA	Forest Optimization Algorithm
9	FP	False Positive
10	FSFOA	Feature Selection Using the Forest Optimization Algorithm
11	FSWFOA	Feature Selection Using Enhanced Weighted Forest Optimization Algorithm
12	GA	Genetic Algorithm
13	GBT	Gradient Boosting Trees
14	GR	Gain Ratio
15	GSC	Global Seeding Changes
16	GSO	Global Seeding Operation
17	GWO	Grey Wolf Optimised
18	HIR	Healthcare Representative
19	HR	Human Resources
20	IBM	International Business Machines
21	IG	Information Gain
22	IRF	Improved Random Forest
23	KNN	K Nearest Neighbour
24	LDA	Linear Discriminant Analysis
25	LMT	Logistic Model Trees
26	LR	Logistic Regression
27	LSC	Local Seeding Changes
28	LSO	Local Seeding Operation
29	LSVM	Linear Support Vector Machine
30	LT	Laboratory Technician
31	MD	Manufacturing Director
32	NB	Naive Bayes
33	NMI	Normalized Mutual Information
34	NN	Neural Networks
35	PCA	Principle Component Analysis
36	PSO	Particle Swarm Optimization
37	RD	Research Director
38	RF	Random Forest
39	RS	Research Scientist
40	SE	Sales Executive
41	SR	Sales Representative
42	SU	Symmetrical Uncertainty
43	SVM	Support Vector Machine
44	TP	True Positive
45	WCO	World Cup Optimization
46	XGBoost	Extreme Gradient Boosting