www.revistageintec.net
ISSN: 2237-0722

revista GEINTEC
Gestão, Inovação e Tecnologias

A Comparative Study of Classification of Occupational Stress in the Insurance Sector Using Machine Learning and Filter Feature Selection Techniques

Arshad Hashmi[1]; Waleed Ali[2]; Shazia Tabassum[3]

[1]Assistant Professor, Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Kingdom of Saudi Arabia.

[2]Associate Professor, Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Kingdom of Saudi Arabia.

[3]Associate Professor, International Institute of Professional Studies (IIPS), Ranchi, Jharkhand, India.

Abstract

In recent years, occupational stress mining has become a widely exciting issue in the research field. The primary purpose of this study is to analyze filter feature selection methods for the efficient occupational stress classification model. We propose and examine seven different techniques of filter feature selection such as Chi-Square, Information Gain, Information Gain Ratio, Correlation, Principal Component Analysis, and Relief. The resultant selected features are then used with popular classifiers like Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosted Trees (GBT) for detection of occupational stress in the insurance sector. A survey-based psychological primary occupational stress data set is used to evaluate the relative performance of these methods. This study effectively demonstrated the significance of filter feature selection methods and explained how accurately they could help classify stress levels. This study showed that the Correlation-based feature selection with the SVM classifier obtained the best performance compared to other filter feature selection methods and classification models.

Key-words: Occupational Stress, Feature Selection, Ranking Method, Stress Classification, ANN, SV, RF.

1. Introduction

Stress is the result of the natural reaction of an organism related to internal, external, positive, and negative stimulation. It is found in three different forms alarm, resistance, and exhaustion state [1]. These three factors help in preparing for the flight or fight response to safeguard the body from threats.

Occupational stress [2] results from a mismatch between the demands and working ability of the people. Chronic stress leads to several health issues such as cardiovascular diseases and musculoskeletal disorders, resulting in a drop in job performance [3].

The detection and monitoring of occupational stress in the early stage can reduce health problems and enhance job performance. On the other hand, by recognizing a moderate level of stress appropriate working state can be maintained. Therefore, detecting different levels of stress is meaningful. Stress at work arises from several factors; this includes long work hours, work overload, deadline pressure, high responsibility, lack of training, conflicts, job insecurity, poor physical work conditions, and rotating shift work [4]. Numerous detection approaches have been proposed to identify occupational stress, but very few specifically relate to filter feature selection methods. Most of the existing studies [5, 6, 7, 8, 16, 17, 18, 19, 20, 21, 22] used regression techniques to predict stress. Therefore, effective stress detection and classification modelling technique is required for evaluating the stress of insurance sector personal with reasonable accuracy.

This present paper proposed novel stress classification models using different popular classifiers based on filter feature selection methods. The feature ranking techniques are employed to estimate the relative importance of each feature and assign it a corresponding weight.

The rest structure of the article is organized as follows. Section 2 briefly introduces the existing approaches related to the prediction of stress. Section 3 describes feature selection details used for stress detection. Section 4 presents details of classifiers used along with the feature selection methods. The proposed model is presented in Section 5. Section 6 contains the analysis and discussion of experimental results. Section 7 reports the overall comparison and discussion. Finally, section 8 encloses the conclusion and future work.

2. Related Work

Various studies are carried out to study occupational stress and its related factors, as shown in Table-1. Authors employed regression techniques to detect and analyze stress in most studies [5, 6, 7, 8, 16, 18, 19, 20, 21, 22]. In [5], analysis of variance, including classification and regression tree (CART), has been applied for measurement and modeling of Job Stress. To explore the Effective Factors on Job Stress [6] have applied correlation coefficient test and progressive multivariate regression. In [7], to analyze the occupational stress associated parameters, logistic regression was employed. In [8], ordinary least squares regression is used for identifying stress correlation. Further to the model, the job stress of crane operator regression techniques employed and so on. As shown in Table-1, in most studies, prediction modeling is done using regression techniques. Further, it can also be seen, from Table-1, feature selection techniques have been employed only in very few studies [10, 11] to find the best subset of significant features in recognizing the appropriate features. In order to resolve this problem, machine-learning techniques have been used along with the feature selection methods to discriminate the stress and no stress employee In this study; we will perform a comparative analysis based on seven filter feature selection methods and then performed stress detection modeling using SVM, RF, ANN, NB, and GBDT.

3. Feature Selection

In recent years, in many real-world applications, feature selection and ranking became an active research field and successfully applied with ML techniques [4, 16]. Feature selection is a pre-processing approach to find the subset of the most significant features from the original data in the modeling to manage dimensionality issues. In this strategy, relevant features are considered while irrelevant and redundant are not taken care of. Feature selection algorithms are broadly classified into filters and wrappers depending on the relationship with the learning technique [17]. The filter method extracts features from the data without considering the classifier. In the filter method, feature importance is evaluated by scoring attributes using statistical procedures. It assesses the rank of every individual feature with no consideration of the interrelationship between features. Its main goal is to assign a numerical score to each feature in order to indicate the importance of the feature for classification it may assign higher values for a relevant feature and small values for insignificant ones. We have used the top-k method counts to manage relevant features. As a result, learning performance and classification accuracy are enhanced to some extent. The proposed filter feature selection techniques are briefly described in the following sections.

3.1. Information Gain (IG)

It provides the relevance of the attributes using the concept of entropy. Entropy is a measure of randomness. The most negligible value of entropy is better for classification, and it ranges from zero to one. The IG used the ID3 algorithm as the backbone to assign weights to them accordingly.

3.2. Information Gain Ratio (IGR)

It is the enhanced format of IG to compensate for its bias. It biases against considering attributes with a large number of distinct values. Although, Information Gain is often a good measure to determine the degree of importance of a feature**.**

3.3. Correlation (CR)

The correlation approach assesses how well an individual feature contributes to the separation of class. CR value lies between -1 and +1. A higher correlation value is considered better for discriminating the features for classification.

3.4. Chi-Squared Statistic (X2)

It computes the weight of individual features and assigns a ranking score. The relevance of feature is based on the highest ranking. Its value is given by

$$\text{X2}= \sum[(OF-EF)2/EF \qquad (1)$$

X2 is the chi-square statistic, OF is the observed frequency, and EF is the expected frequency.

3.5. Principal Component Analysis (PCA)

The PCA generates attribute weights of the given an example set employing a covariance matrix. The higher the importance of an attribute, the more relevant it is considered [23].

3.6. Relief Method

The relief method employed sampling technique to find the relevant feature and then compared the value of the recent feature for the nearby instance of the similar and a dissimilar class. It depends on the nearest neighbors. It picks the feature that is the most differentiable among the various classes [24] based on highest relevant scores.

3.7. Gini Index (GI)

It is used to quantify the distribution of the feature concerning classes. The impeccability introduces the separation level of a feature to recognize the potential classes [25]. For a feature, the GI is determined by the condition.

$$GI(t_i) = \sum_{j=1}^{m} p(\mathrm{ti|Cj})^2\, p(\mathrm{Cj|ti})^2 \qquad (2)$$

Where m is the number of classes, $p(\mathrm{ti|Cj})$ is the term ti probability of the given class Cj, $p(\mathrm{Cj|ti})$ is the class Cj probability of the given the term ti.

Table 1 - Summary of Existing Work Related to Occupational Stress Prediction

Approach	**Machine Learning Techniques**
Intelligent modeling of job stress of electric overhead traveling crane operators[5]	CART
Effective Factors on Job Stress from Experts' Perception; a Case Study in Iranian Agriculture Engineering Organization[6]	Multivariate regression
An efficient evaluation of prevalence and associated parameters of occupational stress [7]	Logistic Regression
Intelligent comparison for associating correlates of stress [8]	OLS Regression
An intelligent stress-detection system[9]	Fuzzy Logic
Evaluating feature selection for stress identification[10] Linear Discriminant Function-- LDF	LDF, Induction tree, SVM,NB and KNN
Assessing Job stress using response surface data mining[11]	Response Surface Methodology
Automated stress evaluation based on EEG signal[12]	ANN,SVM,LDA,
Intelligent techniques for Stress Recognition in Reading[13]	ANN, GAs, SVM
ANN based occupational stress modeling[14]	ANN
Occupational stress Prediction modeling on generic basis[15]	Bayesian Networks
Measurement and modeling of job stress of electric overhead traveling crane operators [16]	Regression
Occupational stress and job demand, control and support factors among construction project consultants [17]	Regression
Predicting Occupational Stress for Women Working in the Bank with Assessment of Their Organizational Commitment and Personality Type.[18]	Regression
An investigation on occupational stress of the operating room staffs in hospitals affiliated to Isfahan University of Medical Sciences and its association with some factors[19]	Regression
The impact of role stress fit and self-esteem on the job attitudes of IT professionals[20]	Regression
Job stress and coping strategies in health care professionals working with cancer patients[21]	Regression

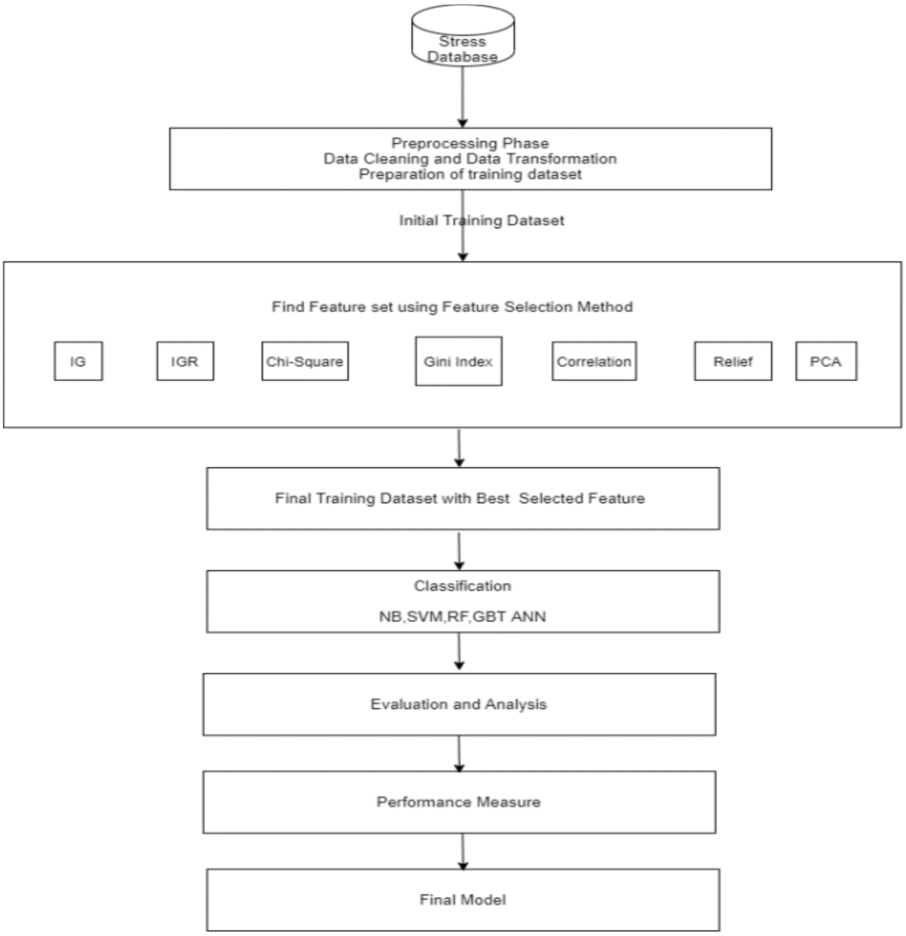


Figure 1 - Framework of the Study

4. Classification Algorithms

The five most popular classification algorithms used in this study are SVM, ANN, NB, Random Forest, and GBDT. Machine-learning techniques work to build classification models in two phases named as training and testing phase. In the training phase, a model is developed from a set of training data with the target outputs, while the testing phase estimates the quality of the trained models from the testing dataset without the expected outcome.

4.1. Support Vector Machine (SVM)

SVM algorithm by means of hyperplane separates the dataset into classes. The test data belong to which class is decided by the hyperplane [26]. Many hyperplanes can exist and based on max-margin between data points the best hyperplane is selected .The dataset nearest to the hyperplane is called a support vector.

Received: 20.09.2021 – Accepted: 12.10.2021

Consider a set of training data vectors

$$xx = \{xx1, xx2, \ldots \ldots \ldots \ldots \ldots xxnn\}, xxii€RRdd \qquad (3)$$

and a set of corresponding labels

$$YY = \{yy1, yy2, \ldots .yynn\}, yyii€\{1, -1\} \qquad (4)$$

To find an ideal hyperplane, SVM maximizes the margin between the separating hyperplane and the closest instance in each class. The hyperplane can be expressed as in Eq.

$$(z, a) + c = 0\ w€\ Rd, c€R \qquad (5)$$

where the vector z defines the boundary, a is the input vector of dimension d, and c is a scalar threshold.

4.2. Random Forest (RF)

An ensemble classifier assesses numerous decision trees and totals their outcomes following majority votes [27]. There is two-level randomization in building these models. In the first place, the bootstrap sample is considered the training data, and then each tree is trained on it. After that, recursive iteration is performed in the next phase to construct a decision tree, and a random selection of feature subset is utilized for further evaluation. In this exploration, we developed and assessed Random Forest (RF) with ten trees. Trees and the results included in Random Forest are based on the majority of accurate output.

4.3. ANN

ANN is the most popular classifier based on the simulation of biological neurons. The neurons create the network. The central processing portion of the neuron is the nucleus. The cell receives an input signal employing Dendrites. This input signal is processed by Soma. An axon terminal turns the processed inputs into outputs. The electrochemical contact between the neurons is Synapses, which can enhance or decrease connection strength from neuron to neuron. The ANN architecture consists of input, hidden, and output layers [28]. The hidden layer is responsible for mapping the input layer information with the output layer. The ANN classifier after training can separates the data into stress and no stress category. Backpropagation algorithm is employed for training and the activation function sigmoid is used during the processing. Multilayer Perceptron Model is used to map the set of input data onto a set of suitable output MLP utilizes backpropagation for training the network. The backpropagation algorithm consists of two phases: propagation and weight update. In this line, the correct answer is compared with the output values to compute some predefined error function value. Finally, the network receives the error as fed back. The weights are updated based on the learning rate of the backpropagation algorithm. This process repeated until the error rate converges and the network has learned a specific target function.

4.4. The Naive Bayes (NB)

NB classifier assumes that a specific feature in a class is unrelated to any other feature. NB is a high-bias and low-variance classifier. The beauty is that it can build a model even with a smaller data set. It is computationally inexpensive and based on the Bayesian theorem. The NB employs Gaussian probability densities for modeling purpose.

4.5. Gradient Boosted Decision Trees (GBDT)

This algorithm produces a predictive model in the form of an ensemble by integrating the predictions from multiple decision trees using a boosting approach [29]. The ensemble came into existence in the several stage by gradient descent in function space.

5. The Proposed Intelligent Occupational Stress Detection Model

The framework in **Figure 1** represents the steps from data collection to final model creation. We collect a database from north region of LIC and ICICI using questionnaire and then perform the data cleaning and transformation to make the dataset more consistent. We propose applying seven feature selection methods with the stress dataset. Each feature selection method will select appropriate features that can contribute to enhancing the performance of classification techniques. This study uses five classification methods to classify the stress and no-stress employee from the dataset. The NB, SVM, RF, GBT and ANN will use the feature selected by each feature selection method to classify the stress and no-stress employee. Finally, the performances of classification models will be discussed and analysed.

5.1. Phase of Data Collection

We collected the dataset from 600 working professionals from the LIC and ICICI insurance sector from the northern sector of India by using the questionnaire through online mode and prepare a database. The dataset comprises 56 features divided into 9 demographic features and 46 categorical features in addition to one class feature. Demographic features include Gender, Age, Education, Management, Experience, Spouse, Religion, City, and Company. There are 12 variables related to relevant components that are the root cause of stress in employees. This questionnaire has 46 statements to measure the 12 types of variables used in the study. These variables are Role Overload (RO), Role Ambiguity (RA), Role Conflict (RC), Unreasonable Group and Political Pressure (UGPP), Responsibility for persons (RP), Under-Participation (UP), Powerlessness (PL), Poor Peer Relations (PPR), Intrinsic Impoverishment (IIM), Low Status(LS), Strenuous Working condition (SWC) and Unprofitability (UPF). The class variable contains two values 1 and 0 for stress and non-stress classification, respectively.

5.2. Phase of Preprocessing

To use the dataset in our proposed model we applied preprocessing techniques. First, we removed some of the records having missing values to clean the data. Then integrated the data. After that performed data transformation i.e., converting categorical variable into numeric. Next, we applied normalization and then used data visualization.

5.3. Phase of Feature Selection

In the proposed model, after data gathering and pre-processing 5.2 in the first stage, we have applied feature-ranking methods as discussed in section 3.1 to 3.7. The feature ranking algorithms are applied and their parameters are adjusted to obtain the most relevant features from the dataset. In the second stage, 20 features having highest rank were selected for the final experiment as shown in Table 2 and Table 3. The results shown in Table-2 demonstrate that; Stat-7 gets the highest score by four of the algorithms (Chi-Square, Info Gain, Correlation, and Gini Index). The Info GR gives the highest score to Stat-44. The Relief algorithm gives the highest score to Stat-14 and PCA to the City attribute.

5.4. Phase of Classifiers Training

In this section, we have described how we applied different classifiers on the selected features to observe the model’s stress detection capability related to stress and no-stress factor. We have used 10 fold cross validation. The classification algorithms used were RF, NB, ANN, SVM, and GBDT. The feature selection algorithms are applied and their parameters are adjusted with respect to the classification accuracy in the second stage. We have adjusted the parameters of classification algorithms to use the top 10 to 20 significant attribute for the respective classification algorithms to bring the best classification performance. The Parameters setting was employed for each classification algorithm as shown in Table-3. The accuracy, sensitivity, specificity, and AUC values are recorded for each model for visualization.

6. Analysis and Discussion of Results

In the following sections, we have discussed in brief the training dataset after feature selection and evaluation measures.

6.1. Dataset Collection and Preparation

In the dataset, the target variable is a two-class problem and can be defined as stress and no-stress employee. The dataset included the 500 instances of insurance sector after applying preprocessing. We have 360 instances related to males and 240 related to females, and the number of features was 56. The best 20 features were selected using six different filter feature selection methods to prepare the training dataset for final experiment to train and evaluate the proposed stress detection model using five popular classifiers.

6.2. Evaluation Measures

For the experiment, we have used version 15.0.8 of Rapid Miner. In order to train and evaluate the model, we employed a ten-fold cross-validation technique. Further, we have considered the model efficiency using four performance metrics accuracy, sensitivity, specificity, and area under the ROC curve (AUC). The confusion matrix presented in Table-5 can efficiently explain these measures.

TP denotes correctly classified positive samples i.e. stress employee. TN represents correctly classified negative samples i.e. no stress employee, FP denotes incorrectly classified no-stress employee i.e., classifier detecting no stress employee as stress employee and FN denotes incorrectly classified stress employee i.e. classifier detecting employees facing stress as no stress.

Table 2 - Scores Obtained by Proposed Feature Selection Methods

Wt. by Chi Square		Wt. by Info Gain		Wt. by Info Gain Ratio		Wt. by Correlation		Wt. by Relief		Wt.by Gini Index		Wt. by PCA	
Attribute	Rank	Attribute	Rank	Attribute	Rank	Attribute	Rank	Attribute	Rank	Attribute	Rank	Attribute	Rank
Stat-7	9.643	Stat-7	0.013	Stat-44	0.048	**Stat-7**	0.107	Stat-14	0.082	Stat-7	0.009	City	0.784
Stat-31	9.194	Stat-31	0.012	Stat-35	0.029	**Stat-17**	0.095	Stat-46	0.079	Stat-31	0.008	Stat-10	0.136
Stat-12	8.614	Stat-38	0.010	Stat-7	0.026	**Stat-38**	0.083	Stat-34	0.076	Stat-38	0.007	Stat-11	0.130
Stat-38	7.735	Stat-34	0.008	Stat-41	0.025	**Stat-42**	0.081	Stat-45	0.053	Stat-34	0.006	Stat-24	0.115
Stat- 2	7.337	Stat-44	0.008	Stat-31	0.023	**Stat-31**	0.081	Stat-16	0.052	Stat-44	0.005	Stat-15	0.104
Stat-34	6.167	Stat-12	0.007	Stat-6	0.020	**Stat-34**	0.076	Stat-17	0.051	Stat-12	0.005	Stat-37	0.100
City	6.162	Spouse	0.006	Stat-12	0.017	**Stat-20**	0.070	Stat-18	0.048	Spouse	0.004	Stat-46	0.094
Stat-44	6.067	Stat-17	0.006	Stat-17	0.017	**Stat-39**	0.070	Stat-27	0.044	Stat-17	0.004	Stat-25	0.092
Stat-46	6.012	Stat-6	0.005	Spouse	0.016	**Age**	0.067	Stat-31	0.044	Stat-6	0.004	Stat-3	0.085
Stat-24	5.719	Age	0.005	Stat- 1	0.015	**Stat-11**	0.067	Stat-42	0.040	Age	0.004	Stat-26	0.083
Stat-17	5.606	Stat-43	0.005	Stat-34	0.015	**Stat-44**	0.066	Stat-44	0.040	Stat-43	0.003	Stat-36	0.082
Stat-41	4.962	Stat-46	0.005	Stat- 2	0.014	**Stat-45**	0.065	Stat-24	0.037	Stat-46	0.003	Stat-19	0.078
Stat-43	4.804	Stat-42	0.005	Stat-42	0.014	**Stat-35**	0.064	Stat-41	0.037	Stat-42	0.003	Stat-42	0.062
Stat-11	4.776	Stat-39	0.005	Stat-38	0.011	**Stat-46**	0.061	Stat-3	0.035	Stat-39	0.003	Stat-8	0.061
Stat-3	4.731	Stat- 2	0.005	Stat-46	0.010	**Spouse**	0.059	Stat-38	0.034	Stat- 2	0.003	Stat-44	0.056
Stat-42	4.549	Stat-41	0.005	Stat-39	0.009	Stat-12	0.052	Stat-12	0.032	Stat-41	0.003	Religion	0.056
Spouse	4.401	Stat-11	0.004	Stat-16	0.008	Stat-26	0.052	Stat-8	0.029	Stat-11	0.003	Stat- 2	0.054
Stat-4	4.392	Stat-36	0.004	Stat-33	0.008	Stat-21	0.051	Stat-11	0.025	Stat-36	0.003	Stat-34	0.054
Stat-29	4.323	Stat-35	0.003	Stat-29	0.007	Stat-40	0.048	Stat-15	0.017	Stat-24	0.002	Age	0.048
Stat-39	4.257	Stat-22	0.003	Stat-43	0.007	Stat-19	0.047	Stat- 2	0.015	Stat-16	0.002	Stat-12	0.043
Stat-6	4.066	Stat-23	0.003	Stat-32	0.007	Gender	0.046	Stat-7	0.014	Stat-35	0.002	Stat-7	0.032
Age	4.032	Stat-16	0.003	Stat-24	0.007	Stat-30	0.042	Stat-22	0.013	Stat-29	0.002	Experience	0.032

Table 3 - Attributes Details Selected by Correlation Feature Selection Method

Attribute No	**Description**
Stat-7	Working with person whom I like.
Stat-11	I do my work under tense circumstances.
Stat-17	My cooperation is frequently sought in solving the administrative or industrial problem at higher level.
Stat-20	I get amply opportunity to utilize me abilities and experience independently.
Stat-34	My higher authorities do not give due significance to my post and work.
Stat-35	I often feel that this job has made my life cumber some (ungraceful).
Stat-38	Employees attach due importance to the official instructions and formal working procedures
Stat-39	I am compelled to violate the formal and administrative procedures and policies owing to group/political pressures.
Stat-31	At the place where I work; my opinion seems to be count
Stat-41	There exists sufficient mutual co-operation and team spirit among the employees of the organization/department.
Stat-42	My suggestions and co-operations are not sought in solving even those problems for which I am quite competent.
Stat-44	I have to do such work as ought to be done by other.
Stat-45	It becomes difficult to implement all of a sudden the new dealing procedure and policies in place of those already in practice.
Stat-46	I am unable to carry out my assignment to my satisfaction on account of excessive load of work and lack of time

Table 4 - Parameter Grid

Model	Parameter
RF	No of Trees-100, Dept of Tree-10,Stooping Criteria-Pruning, Voting Strategy-Confidence
ANN	Learning Rate-0.5 Momentum-0.9, No of hidden Layer-1Activation Function-Sigmoid, Epsilon-1.00E-04
SVM	Kernal- Neural, c-0.09
NB	Laplace Correction, Kernel Density estimation-Full, Bandwidth selection-Fix
GBDT	No of Trees-100,Maximal Depth-8, Minimum Rows-10

Table 5 - Confusion Matrix

		Predicted Classification	
Actual Classification		Positive **Stress Employee**	Negative **No Stress Employee**
	Positive **Stress Employee**	True Positive (TP)	False Negative (FN)
	Negative **No Stress Employee**	False Positive (FP)	True Negative (TN)

The first evaluation measure used in this paper is accuracy. It is the overall correctly detected values of stress employees and no-stress employees concerning all employees.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

The second evaluation measure is the sensitivity metrics. It indicates actual stressed employees detected among all working employees. Sensitivity metric is computed using equation

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (7)$$

Specificity specifies the actual non-stress employee (true negatives) that the model can correctly classify. Specificity is computed using the equation

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (8)$$

The last evaluation measure employed in the paper is the Area under the Curve (AUC) [30].

It is a popular method of measuring the suitability of the classifiers to differentiate between stress and no-stress employee. It is used as a summary of the Receiver Operator Characteristics (ROC) curve, which is a two-dimensional diagram. AUC is a trade-off between sensitivity and specificity. If the AUC value is found greater than 0.5, it is considered a good model [31]. We obtained the true positives (stress-employee) and true negatives (non-stress) through this process. Our primary goal in this study was to minimize the false negatives rate i.e., the number of non-stress employees incorrectly identified as stressed employees.

6.3. Experimental Results and Discussion

In this section, the experimental results of five classification models, SVM, ANN, NB, RF, and GBT, using seven different feature selection methods are presented. The five most popular classifiers with filter feature selection methods were trained and compared to find the best model as discussed in section 6.2.

6.3.1. Comparison of Popular Machine Learning Classifiers Performance before Feature Selection

In the following section, we have represented comparative performance of classifiers before feature selection.

Table 6 - Classifier Performance before FS

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	49.80	41.77	**57.38**	0.497
NB	49.60	43.80	55.00	0.498
GBDT	50.20	**62.53**	38.74	0.509
ANN	51.00	51.63	50.35	0.523
SVM	**53.80**	56.63	51.12	**0.558**

Based on Table-6, the result showed that the performance of SVM on stress dataset outperforms in AUC and accuracy measure achieving highest value compared to other classifier. However the GBDT outperformed in sensitivity while RF in specificity.

Thus it can be concluded that the performance classification of SVM is relatively superior compared to most of the others before feature selection.

6.3.2. Comparison of Popular Machine Learning Classifiers Performance after Feature Selection

In the following section, we discussed performance comparison of the classifiers after feature selection in the stress detection. With respect to Table 2, the selected features resulted from the respective feature selection methods were employed in the classifier. For each classifier we adjusted the parameter as shown in Table-4, We changed the range of selected features from 5 to 20 and found the best model with 15 selected features on the basis of classification accuracy and AUC value. Only the best model shown in the following tables for each classifier.

Received: 20.09.2021 – Accepted: 12.10.2021

Table 7 - Comparison of Classifiers Performance after Applying Chi Square Feature Selection Method

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	55.80	52.56	58.91	0.571
NB	**56.00**	53.70	58.12	**0.576**
GBDT	53.60	**83.48**	25.58	0.574
ANN	54.20	54.74	53.91	0.546
SVM	53.20	46.27	**59.66**	0.561

When compared with the Table-5, among the classifiers, we can observe that the classification accuracies and AUC of all five classifiers were significantly improved by applying the chi square-based feature selection method. In terms of sensitivity and specificity, the results in Table-6 showed that most of the machine learning classifiers accomplished better performances after applying chi square based feature selection but the SVM showed a slight drop in the sensitivity while GBDT in specificity value after feature selection. Further, it is clear from Table-7 that the NB outperformed most other classifiers in terms of accuracy (56%) and AUC (0.576) used in this study. In terms of other measure, the GBDT achieved the highest sensitivity 83.40%. On the other hand, the SVM obtained the highest specificity value. The AUC is the trade of between the sensitivity and specificity values. The best AUC achieved by NB indicates that the NB correctly predicted both stress and no-stress employees for all working professionals compared to most of other classifiers used for stress detection modelling.

Table 8 - Comparison of Classifiers Performance after Applying Information Gain Feature Selection Method

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	**57.20**	60.37	54.32	0.568
NB	53.60	45.38	61.15	0.577
GBDT	52.40	**70.23**	35.74	0.520
ANN	53.60	55.51	51.96	**0.579**
SVM	53.80	45.54	**61.63**	0.560

When compared the Table-8 with the Table-6 it can be observed that most of the classifiers showed the enhanced performance in terms of all measure after applying feature selection. It is clear from Table-8, RF outperformed most other classifiers in terms of accuracy with highest value 57.20%. While GBDT obtained highest sensitivity value, 70.23% and SVM achieved highest specificity value 61.63%. But ANN, on the other hand, outperformed most other classifiers in terms of AUC. This indicates that ANN can be able to correctly classify both stress and no-stress employee in comparison to most other classifiers.

Table 9 - Comparison of Classifiers Performance after Applying IG Ratio FS Method

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	50.20	48.57	51.48	0.534
NB	**52.20**	44.97	**58.91**	**0.567**
GBDT	49.20	**67.87**	31.83	0.526
ANN	52.20	48.75	55.42	0.520
SVM	52.20	45.45	58.46	0.549

When compared the performance measure with Table-6, it is observed that RF, NB, and ANN showed overall enhanced performance after applying feature selection while GBDT and SVM showed a slight drop in accuracy value. Further, it is clear from Table-9 GBDT achieved the highest sensitivity, 67.87%, but NB outperformed most other classifiers in terms of accuracy 52.20%, specificity 58.91%, and AUC as 0.567. This indicates that NB could correctly classify both stress and no-stress employee compared to most other classifiers.

Table 10 - Comparison of Classifiers Performance after Applying Co-relation Feature Selection Method

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	55.60	49.67	61.26	0.571
NB	54.80	52.56	57.12	0.603
GBDT	51.00	**68.58**	34.49	0.544
ANN	54.80	51.70	57.79	0.572
SVM	**58.23**	56.65	**62.05**	**0.604**

When compared the performance measure with Table-6, it is observed that RF, SVM, and ANN showed overall enhanced performance after applying feature selection while GBDT showed a slight drop in the sensitivity. It is clear from Table-10; GBDT obtained the highest sensitivity, 68.58%. Further, the SVM outperformed most other classifiers in terms of accuracy, specificity, and AUC. This indicates that SVM correctly classified both stress and no-stress employee compared to most other classifiers.

Table 11 - Comparison of Classifiers Performance after Applying Relief Feature Selection Method

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	55.00	50.90	**58.92**	0.512
NB	50.60	47.76	53.14	0.529
GBDT	51.60	**68.62**	35.72	0.532
ANN	**55.00**	52.12	57.72	**0.567**
SVM	49.40	40.93	57.40	0.525

When compared the performance measure of Table-11 with Table-6, it is observed that except SVM, all four classifiers showed enhanced performance in terms of accuracy, sensitivity, and AUC after applying feature selection, while SVM showed a slight drop in accuracy and sensitivity. It is clear from Table-11; in terms of sensitivity, GBDT achieved the highest value of 68.62%, while RF achieved the highest specificity value of 58.92%. On the other hand, ANN outperformed most other classifiers in terms of accuracy and AUC. This indicates that ANN could correctly classify both stress and no-stress employee compared to most other classifiers. After making the comparison of Table-12 with Table-6, it is observed that RF, NB, GBDT, and ANN showed overall enhanced performance in all the measures while SVM showed a little drop in accuracy and sensitivity while GBDT in specificity value after applying feature selection. It is evident from Table-12 that NB achieved highest accuracy and AUC while RF achieved highest specificity. NB outperformed most other classifiers in terms of accuracy and AUC. This indicates that NB was able to correctly classify both stress and no-stress employee compared to most other classifiers.

Table 12 - Comparison of Classifiers Performance after Applying Gini Index Feature Selection Method

Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	50.60	46.63	**58.92**	0.511
NB	**53.20**	47.50	58.52	**0.592**
GBDT	52.40	**67.72**	37.89	0.538
ANN	52.60	53.40	52.05	0.549
SVM	53.20	47.95	58.17	0.554

Table 13 - Comparison of Classifiers Performance after Applying PCA Feature Selection Method

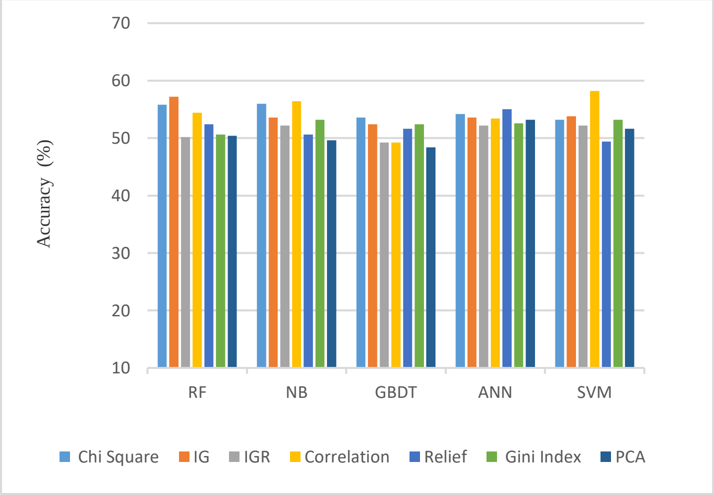
Classifier	Accuracy	Sensitivity	Specificity	AUC
RF	50.40	44.23	56.25	0.499
NB	49.60	49.62	49.72	0.481
GBDT	48.40	**64.35**	33.32	0.498
ANN	**53.20**	55.10	51.63	**0.540**
SVM	51.60	44.58	**58.21**	0.510

The performance comparison of Table-6 with Table-13 revealed that most of the classifiers showed overall poor performance after applying feature selection. However, it is clear from Table-13 that SVM achieved the highest specificity, 58.21%, while GBDT achieved the highest sensitivity, 64.35%. However, the ANN outperformed most other classifiers in terms of accuracy and AUC. This indicates that ANN was able to correctly classify both stress and no-stress employee compared to most other classifiers.

7. Overall Comparison and Discussion

In this section, the overall comparison of accuracy, sensitivity, specificity, and AUC after applying feature selection is presented and discussed.

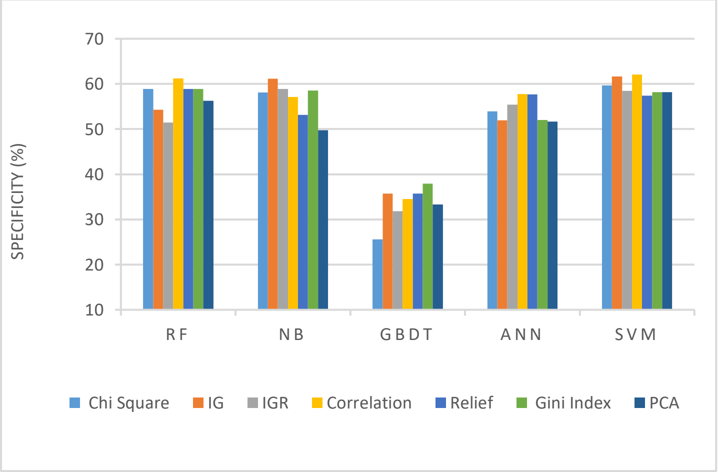
Fig. 2 - Comparison of Accuracy of Classifiers after Applying Feature Selection



It is clear from Fig.2 that SVM with Co-relation based feature selection achieved the best accuracy value. This indicates that the co-relation feature selection method contributed to enhancing the performance in terms of the classification accuracy in detecting the stress and no-stress.

In particular, it is evident from Fig.2 that SVM with correlation methods achieved the best classification accuracy, 58.23%, while GBDT with PCA method attained the weakest classification accuracy, 48.40%, among other learning classifiers.

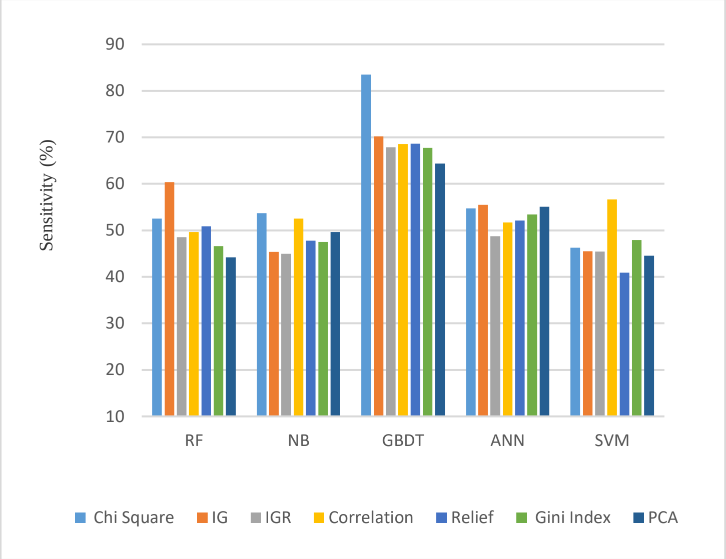
Fig. 3 - Comparison of Specificity of classifiers after applying feature selection.



ISSN: 2237-0722
Vol. 11 No. 4 (2021)
Received: 20.09.2021 – Accepted: 12.10.2021

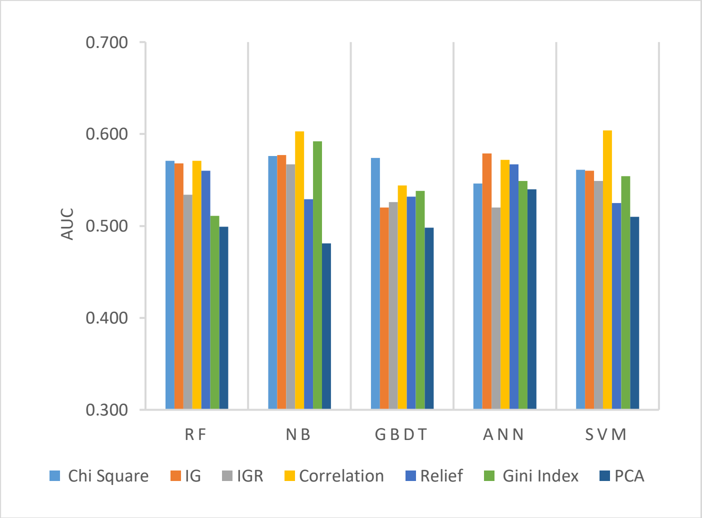
This figure demonstrates that the SVM with co-relation based feature selection method achieved the highest specificity, 62.05%, while GBDT with Chi-Square feature selection achieved the weakest specificity, 25.58%. It is clear from Fig.3 that co-relation based feature selection contributed to enhance the performance of SVM in terms of specificity.

Fig. 4 - Comparison of Sensitivity of Classifiers after Applying Feature Selection



It is evident from Fig.4 that GBDT with Chi Square-based feature selection achieved the highest sensitivity, 83.48%. In contrast, SVM with relief-based feature selection showed the least sensitivity, 40.93%, compared to most other classifiers. This indicates that Chi square FS method contributed to enhance performance measure of GBDT in terms of sensitivity.

Fig. 5 - Comparison of AUC of Classifiers after Applying Feature Selection



It is evident that SVM achieved the best AUC value of 0.604 with co-relation feature selection while NB obtained the least AUC value of 0.481 with the PCA feature selection method. This indicates that co-relation contributed to enhance the AUC value.

Further, for the SVM classifier with co-relation feature selection AUC achievement, the selected features rank and features are shown in Table-2 and Table-3, respectively. This implies that the features chosen by co-relation are more helpful in constructing an accurate model, which can provide a reasonable basis for further screening of stress and no-stress personal in the LIC and ICICI insurance sectors.

AUC is a trade-off between sensitivity and specificity. The best stress detection system must have balanced performance between sensitivity and specificity.

The highest accuracy, specificity, and AUC were accomplished by SVM with co-relation, but the sensitivity is relatively low (Table-10). However, it is observed sensitivity, and specificity value somewhat showed a balanced performance. Because of this, SVM produced balanced detection between stress and no-stress employee.

The main goal of feature selection is to find the best minimal feature subsets to distinguish the stress and no-stress employee. The results reported in Fig.5 showed SVM model with co-relation feature selection achieved this goal.

8. Conclusion and Future Work

In this work, a filter based feature selection methods have been applied to find potential attributes for identifying stress and no-stress personal from the LIC and ICICI insurance domain. Then, the best features have been utilized to train the most popular machine learning techniques in order to identifying stress and no-stress.

The SVM machine learning classifiers considering correlation-based feature selection using 15 attributes outperformed the performances of other classifiers with applying other feature selection. Thus, we can conclude that correlation feature selection with SVM is a suitable technique for classifying stress and non-stress personal and will provide an appropriate mechanism for identifying stressed employees.

In future work, we will apply other feature selection techniques like wrapper method, evolutionary method and hybrid feature selection to enhance accuracy.

References

Office for National Statistics, Social and Vital Statistics Division and Northern Ireland Statistics and Research Agency. Central Survey Unit, Labour Force Survey 19752010. Colchester, Essex: UK Data Archive, 2010.

S.L. Sauter, L.R. Murphy, J.J. Hurrell, Prevention of work-related psychological disorders: a national strategy proposed by the National Institute for Occupational Safety and Health (NIOSH), *Am. Psychol. 45*(10), 1990, 1146.

G. Nema, Y.M. Dhanashree Nagar, A study on the causes of work related stress among the college teachers, *Pac. Bus. Rev. Int.,* 2010, 1–7.

Virtanen, M., Kurvinen, T., Terho, K., Oksanen, T., Peltonen, R., Vahtera, J., Routamaa, M., Elovainio, M., Kivimaki, M., 2009. Work hours, work stress, and collaboration among ward staff in relation to risk of hospital-associated infection among patients. *Med. Care* 47 (3), 310–318.

Sun, X., Liu, Y., Xu, M., et al.: 'Feature selection using dynamic weights for classification', *Knowl.-Based Syst.,* 2013, 37, pp. 541–549

O.B. Krishna, J. Maiti, P.K. Ray, B. Samanta, S. Mandal, and S. Sarkar, "Measurement And Modeling of Job Stress of Electric Overhead *Traveling Crane Operators*," *Saf. Health Work*, Vol. 6, Issue. 4, pp. 279–288, 2015.

Arezou Khaleghi, Maryam Omidi Najafabadi, Frahad Lashgarara, "Effective Factors on Job Stress from Experts' Perception; a Case Study in Iranian Agriculture Engineering Organization", International *Journal of Review in Life Sciences,* Vol. 5, Issue 1, pp. 94-99, 2015.

M. Lotfizadeh, B. Moazen, E. Habibi, and N. Hassim, "*Occupational Stress among Male Employees of Esfahan Steel Company, Iran: Prevalence and Associated Factors*," *Int. J. Prev. Med.,* 4(7), 803-808, 2013.

G.S. Armstrong and M.L. Griffin, "*Does The Job Matter? Comparing Correlates of Stress among Treatment and Correctional Staff in Prisons*," *J. Crim. Justice*, Vol. 32, Issue. 6, pp. 577–592, 2004.

A. De Santos Sierra, C. Sánchez Ávila, J. Guerra Casanova, and G. Bailador Del Pozo, "*A Stress-Detection System Based On Physiological Signals and Fuzzy Logic*," *IEEE Trans.*, 58(10), 4857-4865, 2011.

Y. Deng, Z. Wu, C. H. Chu, and T. Yang, "Evaluating Feature Selection for Stress Identification," *IEEE Conf.*, 584–591, 2012

Y. Lee and S. Shin, "Job Stress Evaluation Using Response Surface Data Mining," *Int. J. Ind. Ergon.*, Vol. 40, Issue. 4, pp. 379–385, 2010

N. Sharma and T. Gedeon, "*For Stress Recognition in Reading*," pp. 117–128, 2013

B. Alić, D. Sejdinović, L. Gurbeta, and A. Badnjevic, "Classification of Stress Recognition Using Artificial Neural Network," *Conf. Embed. Comput.*, 297–300, 2016.

S.G. Herrero, M.A.M. Saldana, J. G. Rodriguez, and D.O. Ritzel, "Influence of Task Demands on Occupational Stress: Gender Differences," *J. Safety Res.*, Vol. 43, Issue. 5, pp. 365–374, 2012

A. De Santos Sierra, C. Sánchez Ávila, J. Guerra Casanova, and G. Bailador Del Pozo, "A Stress-Detection System Based on Physiological Signals and Fuzzy Logic," *IEEE Trans. Ind. Electron.*, 58(10), 4857–4865, 2011

O.B. Krishna, J. Maiti, P.K. Ray, B. Samanta, S. Mandal, and S. Sarkar, “Measurement and Modeling of Job Stress of Electric Overhead Traveling Crane Operators,” *Saf. Health Work*, 6(4), 279–288, 2015.

A. Khaleghi, M.O. Najafabadi, and F. Lashgarara, *“Effective Factors on Job Stress from Experts’* Perception; *A Case Study In Iranian Agriculture Engineering Organization,”* 5(1), 94–99, 2015

P. Bowen, P. Edwards, H. Lingard, and K. Cattell, “Occupational Stress and Job Demand, Control and Support Factors Among Construction Project Consultants,” *Int. J. Proj. Manag.,* 32(7), 1273–1284, 2014

M. Khodabakhshi, “Predicting Occupational Stress for Women Working in the Bank with Assessment of Their Organizational Commitment and Personality Type,” *Procedia - Soc. Behav. Sci.,* 84, 1859-1863, 2013.

S. Bakhtiari, T. Mehrabi, and A. Hasanzadeh, *An Investigation on Occupational Stress of the Operating Room Staffs in Hospitals Affiliated to Isfahan University of Medical.*

Received: 20.09.2021 – Accepted: 12.10.2021