Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432259

Employee Attrition Using Machine Learning And Depression Analysis

Mr Richard Joseph
Asst Prof., Dept. Of Computer Engineering
VESIT
Mumbai, India
richard.joseph@ves.ac.in

Mr Shreyas Udupa
Dept. Of Computer Engineering
VESIT
Mumbai, India
2018.shreyas.v@ves.ac.in

Mr Sanket Jangale
Dept. Of Computer Engineering
VESIT
Mumbai, India
2018.sanket.jangale@ves.ac.in

Mr Kunal Kotkar
Dept. Of Computer Engineering
VESIT
Mumbai, India
2018.kunal.kotkar@ves.ac.in

Mr Parthesh Pawar
Dept. Of Computer Engineering
VESIT
Mumbai, India
2018.parthesh.pawar@ves.ac.in

***Abstract*—Amongst the significant issues that corporate leaders have to deal with within an organization is the decline in proficient employees. This decline is primarily attributed to extreme work pressure, dissatisfaction at work, and ignored mental health issues such as depression, anxiety, etc. This is known as Employee Attrition or Churn Rate. Given the amount of stress employed people go through, focus on the state of mind has gained much- needed traction. Our model aims to predict the employee attrition rate and the employees' emotional assessment in an organization. A survey containing attrition-related questions helped us gather the required data for analysis. Our model will predict the attrition and give the depression analysis with the help of this data. Algorithms such as Decision Tree Classifier (DTC), Support Vector Machine(SVM) and Random Forest Classifier(RFC) were applied to this dataset after performing preprocessing steps, which helped us achieve an accuracy of 86.0% in predicting attrition rate. The results have been expressed using the primary classification metrics, including F1-score and accuracy.**

***Index Terms*—Attrition, Depression, Support Vector Machine, Random Forest**

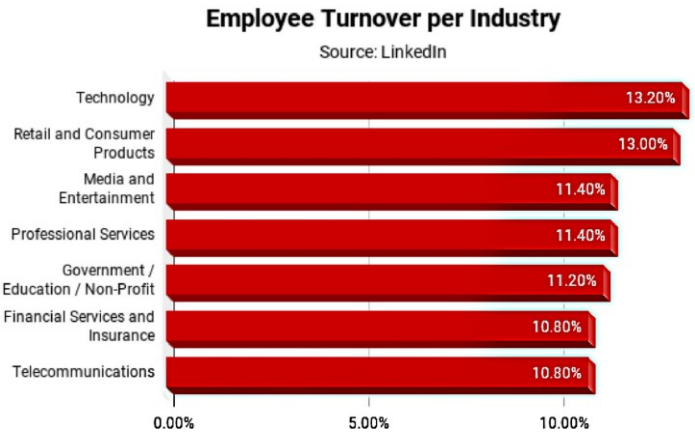

Fig. 1. Attrition Trends in industries

I. INTRODUCTION

Employee turnover [1] can be described as a constant decline in the workforce due to retirement, death, or resignation. Every organization needs to have a certain percentage of attrition to ensure the growth of the organization. Positive attrition is considered beneficial as it generally results in incapable and less productive employees quitting the organization. Meagre attrition rates result in the stagnation of ideas in the workplace. They do not promote intellectual growth caused by exposure to new fresh recruits' new ideas. High attrition rates prove to be exorbitant for the corporation as the corporation invests time, money, and assets to train employees to make them prepared for the job in a particular corporation. In the case where employees quit the job, it causes considerable losses to the corporation. Companies have an uphill task as they must manage recruiting and training recruits and talent loss due to industry attrition trends. Negative attrition implies a larger, more severe problem inside an organization when high-performing employees quit the company searching for better avenues. The losses incurred when an efficient employee quits are not limited to advanced product beliefs, admirable project administration or links with the customers. This can have a detrimental effect on companies as their productivity decreases considerably, which hampers the organization's morale. According to global professional services firm Towers Watson, attrition in India occupies a relatively higher position at 14% compared with global and the Asia Pacific Countries (11.20% and 13.81%, respectively) [2]. Employee churn rate is influenced by several aspects like age, salary, job satisfaction, etc. The elemental takeaway from the considerable employee attrition rate: the corporate world is getting afflicted.

Mental health problems [3] impacts many employees, which is usually disregarded because these problems tend to be hidden at work. The most commonplace mental health disorder that has been studied best in the workplace is depression. Recent studies have shown that employers lose approximately

$44 billion each year due to employees with clinical depression. The constant feeling of sadness and loss of interest in everyday activities, which differs from people's mood fluctuations in daily life, is called depression. A recent study conducted on a survey of a sample reported that about 6%of employees exhibit symptoms of depression in any given year. Besides, employees may be fatigued at work, show signs of presenteeism, absenteeism. Depression may also mar judgment and hampers decision-making. When depression is realistically addressed in the workplace, it promises to lower presenteeism, increase productivity, lower absenteeism and lower medical costs. This paper approves that it is likely to anticipate employee attrition and an employee's mental state in the corporate sector. The prediction, as mentioned above, will help top-level management take preemptive measures to delve into various approaches in retaining their staff, appointing new people or training beforehand. Furthermore, it would assist them to take steps to improve the workplace's mental health scenario.

II. Related Work

Researchers have successfully developed many depression analysis and employee attrition calculation models that can classify expressions, gender, and many other features in recent years.

1. Afef Saidi et al. (2020) [4] presented an innovative audio-based method to detect depression using a hybrid model. Their model combines convolutional neural networks (CNN) and Support Vector Machines (SVM) [13], where SVM is deployed in place of the fully connected layers in CNN. In this proposed model, feature extraction was performed using CNN, and the classification was performed using SVM. They achieved an accuracy of 68% using the hybrid model compared with 58.57% achieved with the CNN model.

2. Akkapon Wongkoblap et al. (2019) [5] collected Facebook users' data from 2007 to 2012 to build a predictive model for detecting symptoms of depression. They used multiple instances of learning neural networks to create their predictive model. This enabled them to develop their model using a few labelled bags instead of requiring all of the labels of the instances used. They achieved maximum accuracy of 74.51% and a precision of 80% in detecting depressed users based on the content on their social network account.

3. Dilip Singh Sisodia et al. (2017) [6] used the HR analytics dataset sourced from Kaggle and tried to build a model that predicts employee churn rate. A correlation matrix and heatmap were generated to show the relation between the attributes. In the experimental section, a histogram was created to compare left employees vs compensation, department, satisfaction level, etc. They used various machine learning algorithms such as Support Vector Machine (SVM), Random Forest Classifier (RFC) [12], K-Nearest Neighbor (KNN) [14], and Na¨ıve Bayes classifier [15] for prediction purposes. Based on this research, we have incorporated the same algorithms for training and testing our model.

4. S. S. Alduayj et al. (2018) [7] used a 3 step experiment process to determine employee attrition rate. In the first experiment, they used an original imbalanced dataset using SVM with various kernel functions KNN, and Random Forest. They concentrated on reducing class imbalance using the adaptive synthetic (ADASYN) approach, retraining the new dataset using the above-mentioned machine learning models. Furthermore, they performed undersampling of the data to achieve a balance between classes. Finally, training an ADASYN-balanced dataset with KNN with K = 3 resulted in the highest performance, with a 0.93 F1-score. They achieved an F!-score of 0.909 while using 12 features out of 29, using Random Forest Classifier and Feature Selection. The essential idea of ADASYN [11] is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn.

5. A carrier company ordered the integration of a revolutionary algorithm named Data Mining Evolutionary algorithm(DMEL) (2003) [8]. Its prime objective was to anticipate the consumer's attrition rate and the chances of them leaving. It was established that if a consumer is leaving, then a set of loyalty programs, including special offers and discounts, are offered by the company to retain the consumer. Applying the model to real-time data showed accurate results by depicting stimulating rules of classification and distinct attrition rates. Using this concept, our model will give the employer/admin suggestions to retain employees on the verge of quitting.

6. M. Deshpande et al.(2017) [9] successfully implemented emotional analysis based on Twitter feeds, primarily focusing on depression. the feed was classified as negative or neutral, based on a specially constructed list of words depicting depression tendencies. They adopted a unique approach to conducting their experiments. By implementing Naive-Bayes Classifier and SVM, they achieved a maximum accuracy of 83.0% in predicting depression tendencies.

7. A study conducted by AR Subhani et al. (2017) [10] found that the human brain is the most affected organ while undergoing stress. This study can be applied to learn the changes and stress that a person with mental illnesses like depression, anxiety, etc., goes through. Several features from the signal analysis of the Electroencephalogram (EEG) on the affected person can be extracted. Classification of the extracted features using algorithms like Decision Tree

Classifier (DTC) [15], Logistic regression etc., providing results that were used to differentiate between stress levels, which helped identify psychological disorders.

III. METHODOLOGY

The following flowchart represents the course of action for the development of our model. (Fig 2)

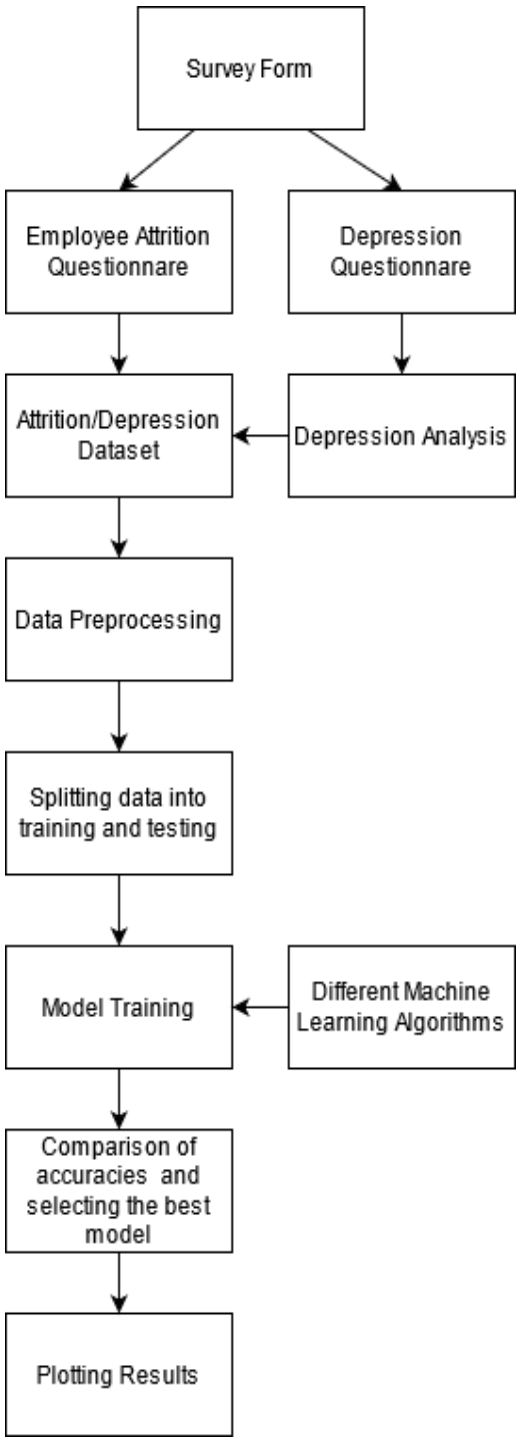

Fig. 2. Flow Of Model

A. Model Training

We have considered some of the features in our dataset, which include work stress, equality at the workplace, job satisfaction, overtime, working hours, travel, personal life, etc. After gathering the data and performing correlation analysis, it was realized that certain features contributed only marginally towards the model's accuracy. Hence, the following attributes, like the stature of the company, canteen facilities, campus environment, etc., were discarded, and the subset mentioned above, was taken into consideration. A significant feature that we have considered in our subset was the analysis performed on the depression questionnaire, containing Goldberg's Depression Questionnaire. Furthermore, the dataset is made to undergo preprocessing to make it suitable for model training. After preprocessing of data, the model undergoes data training where the dataset is split into 75% training and 25% test dataset. Data modelling is performed after the dataset is trained. Different machine learning algorithms such as Support Vector Machines (SVM), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and K-Neighbors (KNN) are used to determine which algorithm results in the best accuracy. To better analyze the results and show the correlation between different features, various graphs have been implemented.

B. Observations

After developing the model on the dataset, we derived the following observations:

- Correlation between the features.
- Dependency of attrition on the features.

Fig 3 shows the relationship between Salary Hike and the Attrition Rate. It can be analyzed that the attrition rate is inversely proportional to the salary hike, i.e. there is a higher chance of attrition when the salary hike is low. This shows that the salary hike is an important feature that affects the employee attrition rate.

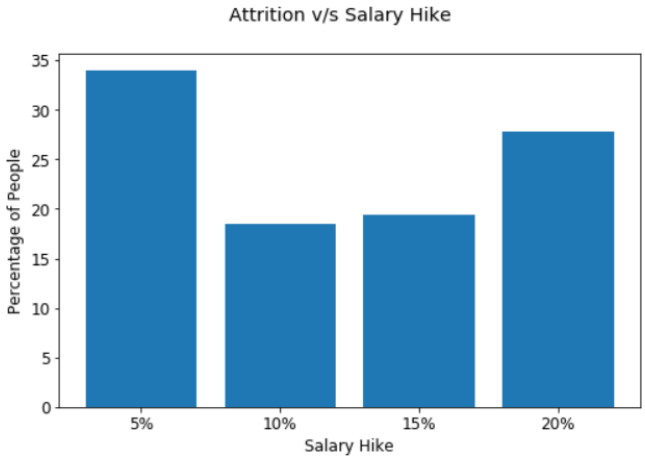

Fig. 3. Attrition v/s Salary Hike

As we can infer from Fig 4, the lower the job satisfaction higher is the percentage of attrition. As the trends suggest, job

satisfaction has a substantial impact on the attrition rate. We can infer from the graph that job satisfaction is an important feature that drives the employee to continue working at the organization.

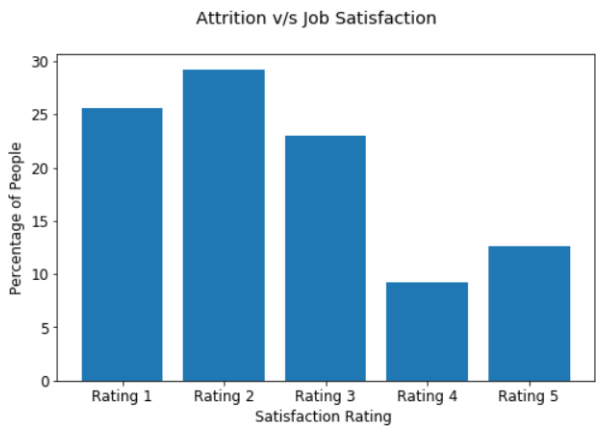

Fig. 4. Attrition v/s Job Satisfaction

From Fig 5, it can be inferred that the two attributes, i.e. equal opportunities for everyone and equal distribution of work, are correlated. The average percentage of attrition when there are equal opportunities and equal work distribution is less than the other two results as expected.

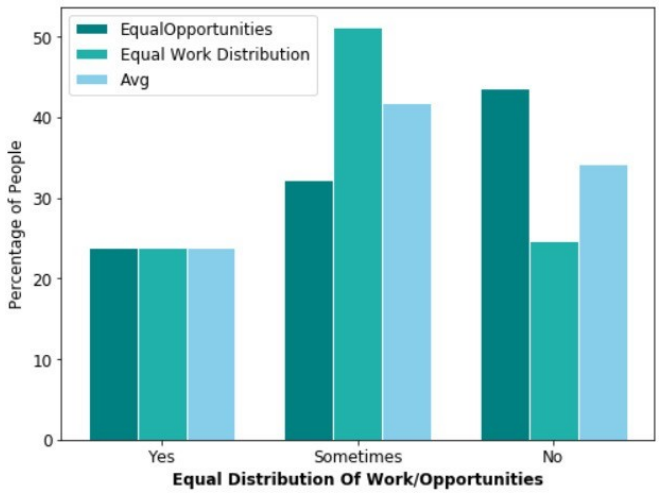

Fig. 5. Correlation between Equal Opportunities and Equal Work Distribution

Fig 6 graph shows a correlation between the balance of professional and personal life and stress at work. For Rating 4 comparatively, the stress is high, and balance is low, so the attrition rate is higher. The average rate of attrition increases with the rating as expected. The graph suggests that as balance in life decreases, the stress increases.

Depression analysis will be done for employees, and the HR Dept will analyze the mental health. Fig. 7 shows the effect of the depression level on attrition. As the depression level goes higher, the chances of attrition are higher. This indicates

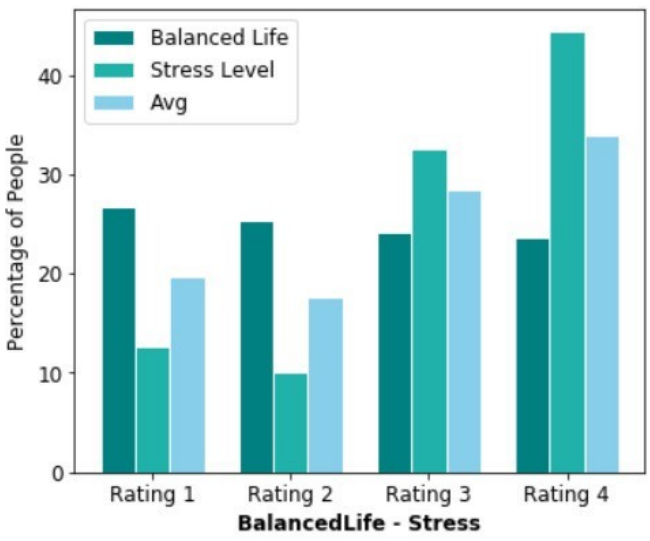

Fig. 6. Correlation between Work-Life Balance and Stress

that mental health is an important measure to be taken into consideration.

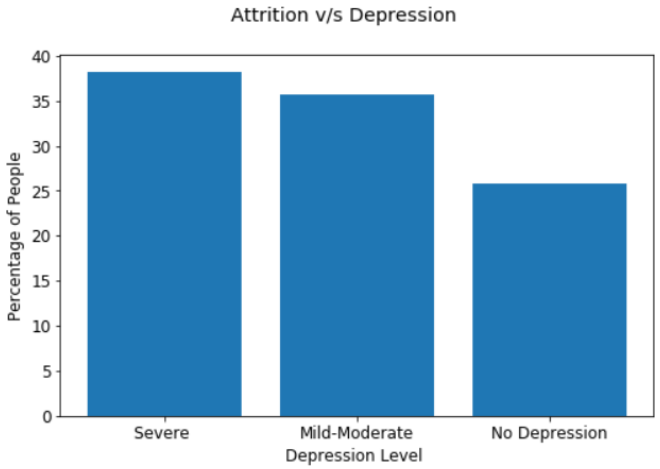

Fig. 7. Attrition v/s Depression

IV. RESULT

This paper demonstrates the use of various classification algorithms Support Vector Machines (SVM), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and K-Neighbors (KNN), for predicting employee attrition rates in an organization.

A comparative study was performed using six different classification algorithms to enhance accuracy. After training all models, the accuracies of the various models were compared. Random Forest Classifier Algorithm tops the list with an accuracy of 86.00%, followed by Gaussian Naive Bayes (GNB) at 81.40%.

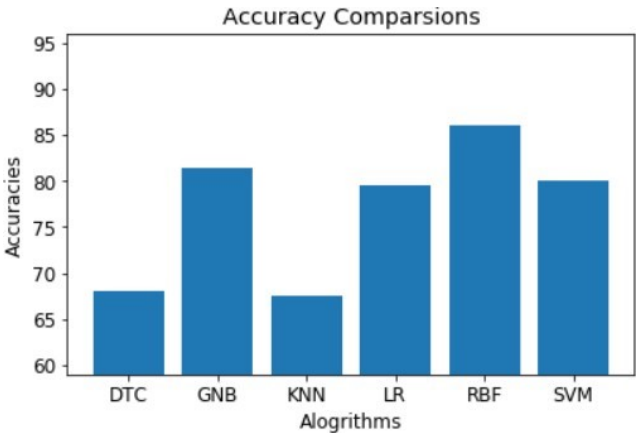

Fig. 8. Comparison of Accuracies

The following fig 8, shows the bar graph representation of accuracies achieved for the different algorithms we have used to analyze the attrition rate. Our study results state that Random Forest Classifier (RFC) has achieved the highest accuracy as the features and the correlation between them is better suited for RFC. The rest of the models exhibited comparatively less accuracy. Table 1 shows the details of the models' accuracy and their mean scores.

TABLE I
COMPARISION OF ACCURACY SCORES

Algorithm	Accuracy	R Square_Score	F1_Score
Random Forest Classifier	86.00%	0.356	0.8599
Gaussian NB	81.40%	0.173	0.814
SVM	80.04%	0.099	0.8004
Logistic Regression	79.60%	0.062	0.796
Decision Tree Classifier	68.00%	-0.470	0.68
KNN	67.6%	-0.488	0.676

V. FUTURE SCOPE

1) All the major companies and govt institutions can make use of our product. This product can be implemented across various sectors like Finance, Education, IT etc. The product can be custom-built according to the different needs of the sectors.
2) The features can be altered to include questions pertaining to Work from Home problems like the ease of access to internet connectivity, personal issues, living space, etc.
3) Emotional analysis can be extended to include other mental health disorders like anxiety, stress etc.

VI. CONCLUSION

Our model considers various features and predicts the attrition and mental health for an individual employee with an accuracy of 86.0%, which is higher than the existing solutions using the same algorithm.

The dataset that we have used has independent as well as correlated attributes. Support Vector Machines (kernel =' poly') is better for non-linear problems but has poor performance when used on a dataset with a large number of features. Furthermore, Naive Bayes(Gaussian) also works well on non-linear datasets with many features, but it needs to have allthe features independent of each other, while there is some correlation in our case. Finally, Random Forest Classification works perfectly for such conditions, where there is a correlation between features and the number of features is large. For the reasons mentioned above, Random Forest Classifier gives higher accuracy than Naive Bayes, followed by SVM. For the remaining algorithms, they need the dataset to be linear, which is not the case, and hence the accuracy for other algorithmsis comparatively low.

Using this model, the employers/HRs can be aware of their employees' mental health and take appropriate steps to prevent the employees' attrition. The HR Dept. can focus on employees that need therapy. With this model's help, business organizations can ensure that their employees work in a positive atmosphere without tainting the business's productivity and efficiency.

REFERENCES

[1] K Sunanda (2017), AN EMPIRICAL STUDY ON EMPLOYEE ATTRITION IN IT INDUSTRIES- WITH SPECIFIC REFERENCE TO WIPRO TECHNOLOGIES Paper 15.pdf (researcher-sworld.com)(Online).
[2] Talapatra, Pradip & Rungta, Saket & Anne, Jagadeesh. (2016). EMPLOYEE ATTRITION AND STRATEGIC RETENTION CHALLENGES IN INDIAN MANUFACTURING INDUSTRIES: A CASE STUDY. VSRD International Journal of Business and Management Research. VI. 251-262.
[3] Mental health problems in the workplace - Harvard Health (Online).
[4] Saidi, S. B. Othman and S. B. Saoud, "Hybrid CNN-SVM classifier for efficient depression detection system," 2020 4th International Conference on Advanced Systems and Emergent Technologies (IC ASET), Hammamet, Tunisia, 2020, pp. 229-234, doi: 10.1109/IC ASET49463.2020.9318302.
[5] Wongkoblap A, Vadillo MA, Curcin V. Modeling Depression Symptoms from Social Network Data through Multiple Instance Learning. AMIA Jt Summits Transl Sci Proc. 2019;2019:44-53. Published 2019 May 6.
[6] D. S. Sisodia, S. Vishwakarma and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 1016-1020, doi: 10.1109/ICICI.2017.8365293.
[7] S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.

[8] Wai, H.A., Chan, K.C.C., Yao, X.: A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. Evol. Comput. 7(6), 532–545 (2003).
[9] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 858-862, doi: 10.1109/ISS1.2017.8389299.
[10] A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel and A. S. Malik, "Machine Learning Framework for the Detection of Mental Stress at Multiple Levels," in IEEE Access, vol. 5, pp. 13545-13556, 2017, doi: 10.1109/ACCESS.2017.2723622.
[11] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.
[12] Parmar, Aakash & Katariya, Rakesh & Patel, Vatsal. (2019). A Review on Random Forest: An Ensemble Classifier. 10.1007/978-3-030-03146-6_86.
[13] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.
[14] Zhang, Shichao & Deng, Zhenyun & Cheng, Debo & Zong, Ming & Zhu, Xiaoshu. (2016). Efficient kNN Classification Algorithm for Big Data. Neurocomputing. 195. 10.1016/j.neucom.2015.08.112.
[15] Berrar, Daniel. (2018). Bayes' Theorem and Naive Bayes Classifier. 10.1016/B978-0-12-809633-8.20473-1.
[16] Patel, Harsh & Prajapati, Purvi. (2018). Study and Analysis of Decision Tree-Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78. 10.26438/ijcse/v6i10.7478.