2023 26th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-5901-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT60459.2023.10441106

Depression Level Classification on Real-Time Data using Machine Learning and Data Mining Techniques

Arpita Das, Tamzid Hasan Nahid, Dola Das and Awal Ahmed Fime
Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Khulna-9203, Bangladesh
Email: arpitadas93332@gmail.com, tamzid1607005@gmail.com,
dola.das@cse.kuet.ac.bd, awalahmedfime984@gmail.com

***Abstract*—Having depression can have devastating effects on one's mental as well as physical health. In addition to contributing greatly to the global burden of disease, depression is a leading cause of disability across the globe. An individual's depression severity should be determined rather than just predicting whether he or she has depression. In this paper, a system to predict the level of depression using different machine learning multi-class classifiers is proposed for real-time dataset. The dataset is collected by asking questions individually which contains 392 data points and 9 features. The features are mainly about the changes in their life related to having depression. After data pre-processing and removing the noise and outliers, the best features are selected using the Step-Forward Feature Selection algorithm. The dataset is split into 80% and 20% for the train and the test set respectively. After this, Random Forest (RF), Logistic Regression (LR) classifier, Bagging Classifier (BC), and Naive Bayes (NB) were constructed using the selected best features. The robustness and performances of the classifiers are compared based on the accuracy, f1-score, and Mean Absolute Error (MAE) values. RF showed an accuracy of around 77%, BC around 70%. LR around 70% and NB around 69%. We found that RF gave better results than other classifiers based on not only its accuracy but also on its f1-score (0.725) which is the highest among other classifiers. RF also has the lowest MAE value (0.271). Finally, the weighted voting ensemble model is used for combining the performance of the classifiers. The accuracy for the ensemble model is 70% and the MAE value is 0.342. This paper is a preliminary report of the proposed system.**

***Index Terms*—Depression level, Feature selection, Random Forest (RF), Bagging Classifier (BC), Logistic Regression (LR), Naive Bayes (NB), Weighted voting ensemble.**

I. INTRODUCTION

The mental health of every individual is a prime concern for everyday life. By 2030, depression can become a big source of worldwide disease burden [1]. Since 2020, the whole world is facing a pandemic situation because of COVID-19. Mainly, the young generation is affected mostly by depression in this pandemic situation. It is shown on a survey that 37% of the members of Gen Z are visiting mental health professionals than any other previous generation for counseling [2]. Depression can make a person even commit suicide. Because of less concern, a person is not even aware of the fact that they are already having depression. So, predicting the presence along with the level of depression is very important in today's life.

Since most of the time, the main focus is given to finding out if a person is depressed or not, it becomes a binary class prediction problem. Most of the previous works mainly tried to solve this binary classification problem [[3],[4],[5]] . But more importance should be given to knowing how depressed a person is. Depression can be predicted using different scales like PHQ-8, BDI-2, GDS, etc., by asking different questions to a person [6]. Since the answers to the questions depict whether a person is in depression or not, the questions should be asked following proper administration and guidelines. Besides, there should be some ways of measuring the level of a person's depression. Several machine learning algorithms can be used to analyze and predict the level of depression using high-frequency datasets [7]. In the paper, the main focus is given to this context by making it a multi-class classification problem.

In the paper, a system to predict the level of depression is proposed by considering a real-time dataset using 392 samples along with 9 features. Here, 7 questions namely sadness, loneliness, changes in sleep patterns, etc. are asked to 392 people about different mental health parameters and changes in their lifestyle patterns along with their age and gender. Each question has four levels from low to high changes. Even they are asked if they are feeling any depression at any time. The levels are (*i*) never, (*ii*) sometimes, (*iii*) most of the time, and (*iv*) all the time. From the collected dataset, the best features for prediction are selected using the Step-Forward Feature Selection algorithm and outliers along with noises are removed from the dataset for better performance. Finally, different machine learning algorithms like Random Forest (RF), Logistic Regression (LR) classifier, Bagging Classifier (BC), and Naive Bayes (NB) are applied to the dataset. These classifiers predict the 4 classes or levels of depression in this work. The performance of each classifier is compared using different parameters and attribute values. By analyzing the results from each of the classifiers, it is found that RF performs better in prediction than the other classifiers used in the system. Finally, the classifiers are combined using a weighted voting ensemble model to enhance the performance of the proposed system.

So our contributions to this paper are:

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

1) Collecting real-time data from different people by asking them questions.

2) Prepossessing collected data to make a numerical dataset having 9 features and depression level with 4 levels.

3) Applying RF, LR, BC, and NB to the dataset for prediction.

4) Combining the classifiers using ensemble model.

The following sections provide more details about the depression level estimation technique. Section II explains some of the works related to this field done previously. Section III explains the flowchart of the whole system, the process of creating the dataset, pre-processing, and the used classifiers for prediction. Section IV analyses and compares the results among the used classifiers and compares the proposed system of the paper with other previous systems. Finally, section V concludes the overall processes of the system and future plan for this work.

II. Literature Review

Innumerable works have been done to counter and predict peoples' mental health. In the paper [3] the system was developed by following the method called Depression, Anxiety and Stress Scale, 21 items (DASS-21) to predict depression, anxiety, and stress. But they only used the dataset for the binary classification of depression, anxiety, and stress. The authors in [4] developed a two-step hybrid machine learning system that uses a dataset of home-based older Chinese adults. They used the Area Under the Receiver Operating Characteristic (AUROC) curve for the evaluation of the prediction outcome. Here, the system is also developed for predicting if a person has depression or not. They also focused on a fixed group of people and all of them were adults.

In paper [8] they followed a machine learning approach to predict mental stress in COVID-19. They made the dataset by taking the response of different people at different levels and data was taken online. But it was only a binary classification that classifies high-level stress or low-level stress during the COVID-19 pandemic period only. Devakunchari et al. [9], they predicted only depression, anxiety, and stress using manually collected data from some questions to people. From the answers from the people, they calculated a score to give a decision about their mental health. In [10] tried to develop a framework using machine learning which can detect depression in humans. They collected the data using video questioning and calculated the score manually and declared the result depending on that score using one threshold value.

Unlike the above ones, the proposed system works to predict depression as a multi-class classification problem. The dataset is created by asking questions about the changes in lifestyle and mental conditions of a person suffering from depression and used for developing and analyzing the level. Instead of calculating scores of the answers of a person, the output of the dataset is created based on the answer to the question asked to the person about how depressed a person feels. The system is a generalized system that can work with people of every age and gender and the system is not based on a particular situation.

III. Proposed Methodology

The proposed system has six stages, i.e., data collection, data conversion, data visualization, data preprocessing, feature selection and model fabrication. Each stage is described below in details and Fig. 1 shows the flow diagram of our proposed system.

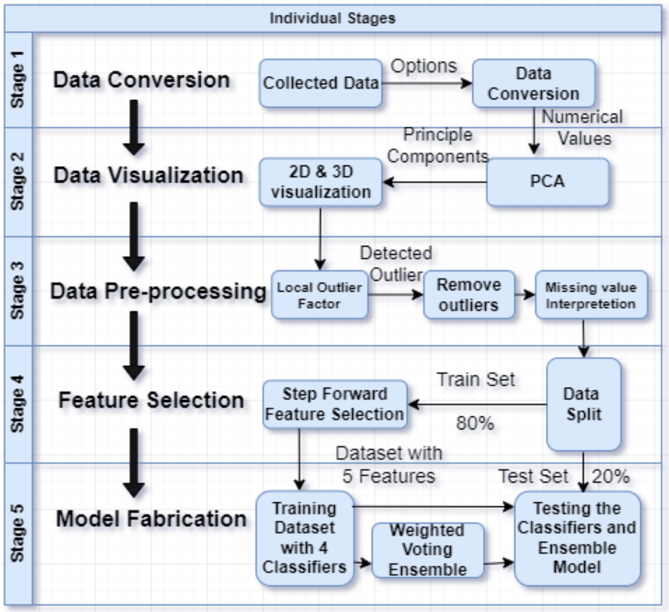

Fig. 1. Flowchart of the proposed system.

A. Data Collection

For data collection, one survey was established and 392 people participated in that survey. People from every occupation and age took part in the survey of which 52.6% were male and 47.4% were female around 18 to 80 ages. Everyone had to answer 10 questions to fulfill the survey. The machine learning models are constructed according to people's own opinions.

The questions used in the survey, are illustrated in Table I.

TABLE I
ATTRIBUTES OF COLLECTED DATA

Attributes
1. Sadness
2. Loneliness
3. Interest in work
4. Energy Level
5. Change in Sleep
6. Change in appetite
7. Tiredness

The answers to these 7 questions were given in a multiple choice of 4 and these are

I) None

II) Low

III) Medium

IV) High

If the participant does not feel any sadness then he will choose none option. If he feels a little sad then he will select low. If he feels much sad then he will select the medium option. If he feels an extreme level of sadness then he will select high. The other questions also have options like sadness.

There are other 2 additional questions about the participant such as

I) Age

II) Gender

The participant must be 18+ to take part in the survey and the upper bound of age is 80. We tried to take data from every profession of men and women.

One last question was set to know about people's actual feelings towards depression. People had to answer by selecting one of the choices among none, low, medium, or high. This question is used as the label output of the model.

B. Data Conversion

The data was collected from various people in string. To make it suitable for the machine learning model, a numerical value was added against every option. As there are 4 options in every question, values 1-4 are added according to their severity.

I) For the None option, 1 is added as it's not severe.

II) For the Low option, 2 is added as it's less severe.

III) For the Medium option, 3 is added as it's moderate-severe.

IV) For the High option, 4 is added as it's so much severe.

C. Data Visualization

After converting the data, Principal Component Analysis (PCA) was used for the visualization of the dataset. Fig. 2 shows the $2D$ visualization of the dataset with two principal components using PCA. Fig. 3 shows the $3D$ visualization of the dataset with three principal components and the total explained variance by the three components is 72.37%. In both $2D$ and $3D$ visualization, In this figure, there are 4 colors for output data points as we have four classes for depression level prediction. Here, the blue color is for the depression level 1 (no depression) class data points, the violet color is for the depression level 2 (low) class data points, the orange is for the depression level 3 (medium) class data points and finally yellow is for the depression level 4 (high) class data points in the dataset.

D. Data Preprocessing

Clustering is performed after dimension reduction to identify noise and outliers in the dataset. Local Outlier Factor (LOF) is one of many clustering methods that is utilized.

Among all the data points, 47 data points were detected as outliers and thus removed from the dataset to increase the performance of the system. The final dataset after deducting the outliers was used for further work.

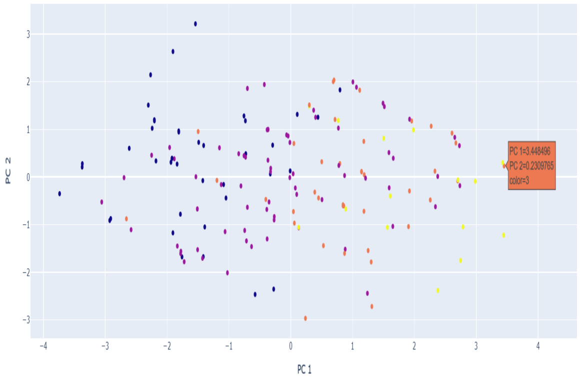

Fig. 2. $2D$ visualization of the dataset using PCA.

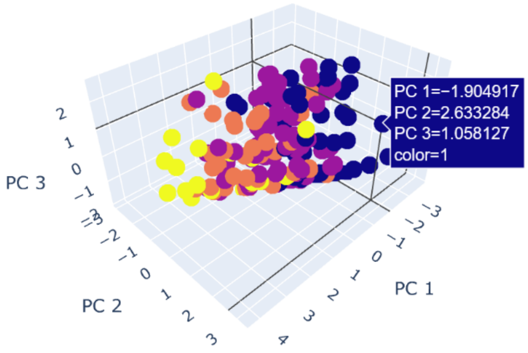

Fig. 3. $3D$ visualization of the dataset using PCA.

Then, if any value is missing in any data point, the missing value is replaced with the highest counted value in that particular column of the dataset.

E. Feature Selection

After detecting outliers and noise and removing them from the dataset, feature selection is done for finding the best features that can increase the performance of the models. The step-forward feature selection method is used for selecting the best features for prediction.

The following attributes are selected as best features in the system by using step-forward feature selection method:

1) Sadness

2) Loneliness

3) Interest in work

4) Change in appetite

5) Tiredness

The model is trained and obtained accuracy based on the selected features.

F. Model Fabrication

After pre-processing the dataset and visualization, the dataset is split into two sets. 80% data points are used for training and 20% data points are used for testing. RF, BC, LR, and NB models are constructed. For training these models, different hyper-parameters are tuned to fit and give better

TABLE II
RESULTS OF DIFFERENT CLASSIFIERS.

	Precision	Recall	MSE	RMSE	RAE	RRSE
Random Forest	0.765	0.709	0.357	0.598	0.383	0.717
Logistic Regression	0.696	0.627	0.429	0.655	0.484	0.786
Bagging Classifier	0.759	0.647	0.673	0.329	0.386	0.621
Naive Bayes	0.659	0.655	0.40	0.632	0.484	0.759
Ensemble Model	0.752	0.696	0.428	0.654	0.519	0.812

TABLE III
CONFUSION MATRIX FOR RF

		Predicted				
		A	B	C	D	Classified as
Actual	A	7	2	0	0	A = tested for class1
	B	1	26	5	0	B = tested for class2
	C	1	2	18	1	C = tested for class3
	D	0	2	2	3	D = tested for class4

performance. Finally, the weighted voting ensemble model is constructed using the four classifiers mentioned before. A corresponding weight is assigned to each model based on its accuracy score and is provided as input to the ensemble model.

IV. EXPERIMENTAL RESULTS

In the system, the dataset was divided into trainset and testset manually at 80% and 20% respectively. The confusion matrix is calculated for each model with the test data. Here, the confusion matrix has a size of 4×4 as the dataset has 4 prediction classes. And using the confusion matrix, different values, like accuracy, precision, recall as well as f1-score are measured. Then MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), RAE (Relative Absolute Error), and RRSE (Root Relative Square Error) error values are calculated to evaluate the performance of each model. Finally, accuracy, f1-score, and MAE values are used for comparing the prediction performances of these models.

The corresponding results of each model are shown in Table II.

1) Results of random forest: The dataset is split into 80% and 20% for the train and the test set respectively. A random forest (RF) classifier is applied in the training dataset for the prediction of depression level. The confusion matrix for Random Forest has also been displayed in Table III.

Based on the confusion matrix, Receiver Operating Characteristic (ROC) Curve is generated in Fig. 4. The micro-average Area Under the Curve (AUC) value aggregates the contributions of all four classes to compute the average metric. Individual AUC for each class is shown as well. For computing AUC for a particular class, that particular class is considered class 1, and the other three classes are considered class zero combinedly.

TABLE IV
CONFUSION MATRIX FOR LR

		Predicted				
		A	B	C	D	Classified as
Actual	A	5	4	0	0	A = tested for class1
	B	2	24	6	0	B = tested for class2
	C	1	3	17	1	C = tested for class3
	D	0	2	2	3	D = tested for class4

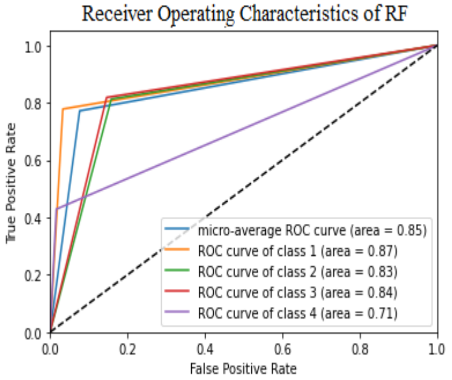

Fig. 4. ROC Curve for Random Forest Classifier.

2) Results of LR classifier: Logistic regression (LR) is also used in the system for the prediction of depression levels. Again, Table IV shows the confusion matrix for LR.

Based on the confusion matrix, Receiver Operating Characteristic (ROC) Curve is generated in Fig. 5.

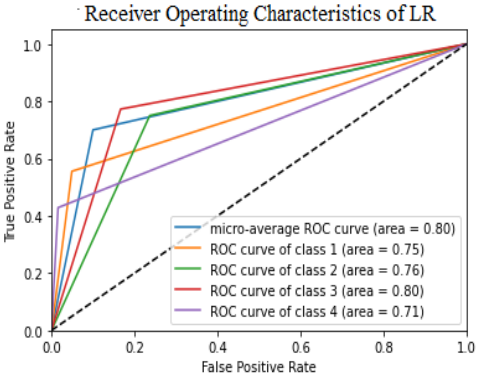

Fig. 5. ROC Curve for LR classifier.

3) Results of Bagging Classifier: Bagging Classifier (BC) is another classifier used in our system for prediction. BC has also given close results like LR. Table V shows the confusion matrix of BC.

Based on the confusion matrix, Receiver Operating Characteristic (ROC) Curve is generated in Fig. 6.

4) Results of Naive Bayes classifier: Naive Bayes (NB) is used as the fourth classifier in the system. Gaussian NB is

TABLE V
CONFUSION MATRIX FOR BC

		Predicted				
		A	B	C	D	Classified as
Actual	A	6	3	0	0	A = tested for class1
	B	2	23	7	0	B = tested for class2
	C	1	4	17	0	C = tested for class3
	D	0	1	3	3	D = tested for class4

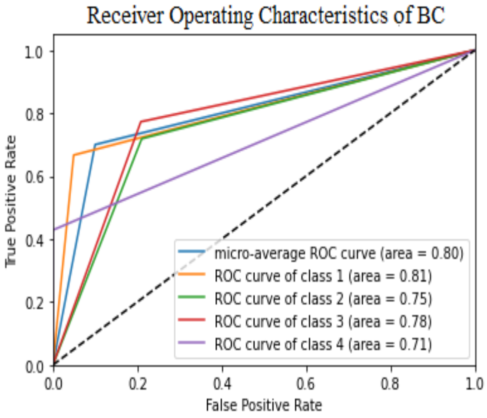

Fig. 6. ROC Curve for Bagging Classifier.

used as it follows the Gaussian distribution for the dataset. The confusion matrix is shown in Table VI.

Based on the confusion matrix, Receiver Operating Characteristic (ROC) Curve is generated in Fig. 6 following the same way as previous classifiers.

A. Results of Weighted Voting Ensemble model

Finally, the four trained classifiers are used as input to the weighted voting ensemble model. The weights are assigned to each classifier according to their accuracy score. The confusion matrix is shown in Table VII.

TABLE VI
CONFUSION MATRIX FOR NB

		Predicted				
		A	B	C	D	Classified as
Actual	A	7	2	0	0	A = tested for class1
	B	3	22	7	0	B = tested for class2
	C	1	3	16	2	C = tested for class3
	D	0	1	3	3	D = tested for class4

TABLE VII
CONFUSION MATRIX FOR WEIGHTED VOTING ENSEMBLE MODEL

		Predicted				
		A	B	C	D	Classified as
Actual	A	7	1	0	0	A = tested for class1
	B	3	25	10	0	B = tested for class2
	C	1	2	14	0	C = tested for class3
	D	0	2	2	3	D = tested for class4

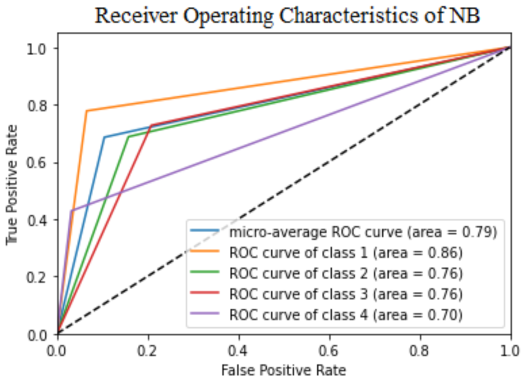

Fig. 7. ROC Curve for Naive Bayes Classifier.

TABLE VIII
PERFORMANCE ANALYSIS OF THE MODELS WITH ACCURACY, F1-SCORE AND MAE

Classifiers	Accuracy(%)	F1-Score	MAE
Random forest	77	0.725	0.271
Bagging classifier	70	0.673	0.329
Logistic regression	70	0.649	0.343
Naive Bayes	69	0.650	0.343
Ensemble Model	70	0.681	0.342

Based on the confusion matrix, Receiver Operating Characteristic (ROC) Curve is generated in Fig. 8 following the same way as previous classifiers.

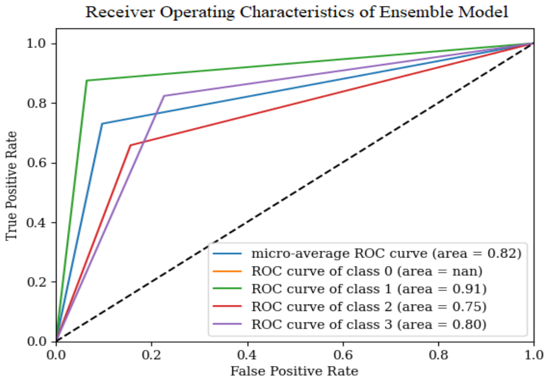

Fig. 8. ROC Curve for Weighted Voting Ensemble Model.

B. Robustness performance of the models

In the system, an analysis of the performance of every model is made using our own dataset of depression-level prediction. For analyzing the performance of each model, the accuracy, F1-score value, and MAE error value are calculated and compared the values for better analysis.

In Table VIII, it is shown that random forest gives the highest accuracy (77%). RF also has a higher F1-score (0.725) and lower error (0.271) than other models. Accuracy can be biased if the class count of the dataset for every class is not the same, but the f1-score is more accurate for uneven datasets.

So, the f1-score is used for analyzing the performance of each model. In Table VIII, bagging classifier and a logistic regression classifier have the same accuracy (70%). But the bagging classifier has a better f1-score (0.673) than the logistic regression classifier (0.650). The bagging classifier has also less MAE value (0.329) than the logistic regression classifier (0.343). So, in the dataset, the bagging classifier is giving better performance than the logistic regression classifier. Finally, the Naive Bayes classifier is giving the lowest accuracy(69%) but the same f1-score (0.650) and MAE value (0.343) as the logistic regression classifier. It means the logistic regression classifier is giving performing better than Naive Bayes based on accuracy. Again, the voting ensemble model gives lower accuracy (70%) than the random forest. Finally, comparatively less accuracy for the models is received because the dataset is created manually in our way by asking different questions to the people. It is mainly a real-time dataset.

In both accuracy and F1-score value comparison, random forest is performing better than the other three models in the system with our dataset for depression level prediction. It also has the lowest error rate among the four classifiers used in the system for prediction.

C. Discussion

In this paper, a dataset is created by asking different questions to different people of different ages and professions. After data pre-processing and selecting the best features, RF, BC, LR, and NB machine learning algorithms are applied to our dataset. Among the 4 models, the random forest classifier performed best for the prediction of the depression level. Finally, a weighted voting ensemble model is generated to combine the prediction of the classifiers based on their accuracy score and give better prediction decisions.

V. Conclusions

Depression affects people's mental health and can drive them in many inadequate directions such as suicidal thoughts. In different situations like the COVID-19 pandemic, people are being triggered to be depressed and their lifestyle changes the most because of depression. Based on this context, an initiative is taken to categorize the level of depression. Prediction of the level of depression is more important because the most depressed person must get more care and support from other people. It is a necessity to know how much depression a person has. In this paper, actual data having 9 features, is collected from all types of people by asking some questions to them during the pandemic situation and lockdown. After completing the dataset having 392 data points, the dataset was pre-processed and collected the best features that can work best for the prediction of depression using the Step Forward feature selection technique. Finally, different Machine Learning classification techniques were applied. RF, LR, BC, and NB models were applied in the construction of the system and the comparison with one another provides the cons and pros of the models. A weighted voting ensemble model is combining the classifier's prediction performances.

In the future, more data points with more features will be added to the dataset for better prediction, and more data pre-processing techniques will be applied to create models with higher accuracy and performance. Data pre-processing steps will be done on only train set after splitting to reduce data leakage. Other mental health-related parameters like anxiety and stress can also be added for prediction with different machine learning algorithms.

References

[1] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar. Classification algorithms based mental health prediction using data mining. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1174–1178, 2020.

[2] Generation z and mental health. https://www.aecf.org/blog/generation-z-and-mental-health, 2023. [Online; accessed 03-March-2021].

[3] E Aidid and R. Musa. Accuracy of supervised machine learning in predicting depression, anxiety and stress using web-based big data: Preserving the humanistic intellect. *Malaysian Journal of Medicine and Health Sciences (eISSN 2636-9346)*, 18(19):87–92, 2022.

[4] Shaowu Lin, Yafei Wu, and Ya Fang. A hybrid machine learning model of depression estimation in home-based older adults: a 7-year follow-up study. *BMC psychiatry*, 22(1):1–13, 2022.

[5] H. M. Abdul Fattah, K. M. Azharul Hasan, and Sunanda Das. A voting classifier for the treatment of employees' mental health disorder. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pages 1–6, 2021.

[6] S. Krishna and J. Anju. Different approaches in depression analysis : A review. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 407–414, 2020.

[7] A. Liu, B. Liu, D. Lee, M. Weissman, J. Posner, J. Cha, and S. Yoo. Machine learning aided prediction of family history of depression. In *2017 New York Scientific Data Summit (NYSDS)*, pages 1–4, 2017.

[8] Luca Flesia, Merylin Monaro, Cristina Mazza, Valentina Fietta, Elena Colicino, Barbara Segatto, and Paolo Roma. Predicting perceived stress related to the covid-19 outbreak through stable psychological traits and machine learning models. *Journal of Clinical Medicine*, 9(10), 2020.

[9] Anu Priya, Shruti Garg, and Neha Prerna Tigga. Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167:1258–1267, 2020.

[10] Ezekiel Victor, Zahra M Aghajan, Amy R Sewart, and Ray Christian. Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological assessment*, 31(8):1019, 2019.