National College of Ireland

A Machine Learning Framework to Predict Depression, Anxiety and Stress

MSc Research Project
Master of Science in Data Analytics

Antony Yesudas
Student ID: x20243405

School of Computing
National College of Ireland

Supervisor: Pramod Pathak,Paul Stynes, Musfira Jilani

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Antony Yesudas
Student ID:	x20243405
Programme:	Master of Science in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Pramod Pathak,Paul Stynes, Musfira Jilani
Submission Due Date:	15/08/2022
Project Title:	A Machine Learning Framework to Predict Depression, Anxiety and Stress
Word Count:	5821
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	17th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Framework to Predict Depression, Anxiety and Stress

Antony Yesudas
x20243405

Abstract

Many people have experienced psychological health problems like stress, anxiety, and depression, over the previous few decades. It is essential to identify mental health conditions quickly and treat them before they worsen. Some people do, however, admit that they struggle with mental illnesses, including stress, anxiety, and depression. Contrarily, the majority of people consider it to be mood changes. Several studies have been done to determine whether user posts on social media can identify mental diseases like depression, anxiety, and stress. This paper predicts psychological issues like stress, anxiety, and depression. Also examined are five distinct severity levels. Random Forest, Naive Bayes and a Neural networks model were used in this research. Data is collected from the publicly available DASS42 tool to apply these algorithms. Random Forest and Naive Bayes performed well for Depression, Anxiety and stress with an average accuracy of 95%, but using evaluation metrics; the Neural networks model was chosen as the best accuracy model with an accuracy of 99%.

Keywords: DASS42, DAS, Machine Learning, Depression, Anxiety, Stress, Gaussian Naive Bayes, Neural Networks.

1 Introduction

Nowadays, people are more ambitious by nature and look for every chance to advance their careers. People now assume that experiencing anxiety, despair, stress, irritation, and unhappiness at work is a normal part of working life. According to the World Health Organization (WHO), depression is the most common mental condition, affecting more than 300 million people globally (Sau; 2017). Because of the severity of the problem, many health researchers have chosen to concentrate their research in this area. Since it is difficult for machines to distinguish between anxiety, depression, and stress, a good learning algorithm is needed for a precise diagnosis. WHO claims that in addition to being physically healthy, a healthy person also has a healthy intellect.

Depression is a mood-based mental illness that manifests as a depressed state of mind, loss of interest, low energy, guilt, trouble sleeping, altered appetite, lack of focus, and hunger. Additionally, it includes anxiety symptoms. This issue may be recurrent or persistent and severely limit the person's capacity to carry out daily tasks. It may potentially result in suicide or self-harm (WFMH; n.d.). Suicide claims the lives of almost one million people annually, with about 3,000 suicide fatalities every day (WHO 2012).

One of the mental diseases that describes how a person's body and mind respond to stress, fear, unease, or unexpected situations is anxiety (Chen X; 2021). Typically, it results in agitation, stress, sweat, and a quick heartbeat. Stress is a state of tension in emotions. Anxiety and stress are important problems nowadays as challenges increase in all areas of life, negatively affecting people's wellness and contributing to real health consequences. It harms a person's academic performance, workplace productivity, and interpersonal connections. Alcohol abuse and a spike in crime rates are other effects on society.

Psychiatrists can establish a patient's DAS level through testing, examinations, and other factors like the patient's medical history. Examples of instruments that can be used to measure DAS include the Hospital Anxiety and Depression Scale (Kaur and Krishnapillai; 2013), the DASS-21 or DASS-42, the Beck Depression Inventory, the Beck Anxiety Inventory (Beck and Garbin; 1988), the Hamilton Rating Scale for Depression, and the CESD (Radloff; 1977).

The 42-question Depression, Anxiety, and Stress Scale checks for these mental illnesses' symptoms. The Patient Health Questionnaire (PHQ) is the accepted method for diagnosing depression at the same time. Because people who experience depression, anxiety, or stress are frequently hesitant to discuss their mental concerns with close friends, family members, or even medical experts, the DASS42 scale was chosen for our research. The dataset's source is the Internet survey data set. It is collected from the online open psychometrics webpage, which contains data compiled from online questionnaires from different people.

1.1 Research Objectives

This research aims to investigate to what extent Can machine learning methods be used to accurately predict mental illnesses like Depression, anxiety, and stress by analysing data relating to individuals' educational backgrounds, lifestyles and other factors. Another goal of this research is to develop a prediction model to detect mental illness early and to know the individual's present condition. Furthermore, I explored the factors like

race, religion, gender, marriage, age groups and academic level of individuals and compared them with their Depression, Anxiety and Stress Conditions.

This paper is formatted as follows: Section 2 examines related studies on stress, anxiety, and depression. While Section 3 outlines the research methodology. Section 4 and Section 5 explain the design specification and implementation procedures. Evaluation and conclusion are in section and section 7, respectively.

2 Related Works

I examined several articles and journals to fully understand the function of machine learning in diagnosing the mental disorder. This section will discuss current approaches to detecting mental illness using machine learning algorithms and how researchers tackle the detection or identification of mental diseases using various strategies. The following paragraphs will detail the research and cutting-edge techniques that were developed.

2.1 Review of previous research based on DAS

(Srinath et al.; 2022), examined how well machine learning algorithms predicted stress, anxiety, and depression levels. The Depression, Anxiety, and Stress Scale, a set of questionnaires, measures the intensity of depression, anxiety, and stress (DASS42). In this work, the performance of Support Vector Machine and Logistic Regression was enhanced, and their classification accuracy was compared to that of other techniques. SVM achieves classification accuracy for depression, anxiety, and stress of 97.35%, 97.49%, and 97.20%, respectively, after parameter adjustment. LR achieves classification accuracy of 98.15%, 98.05%, and 98.45%, respectively, for the Depression, Anxiety, and Stress datasets. The results demonstrated that LR performed better in terms of accuracy than SVM. Their research did not consider deep learning models that potentially reach more accuracy than logistic regression.

(Priya et al.; 2020), made predictions on stress, anxiety, and depression using machine learning algorithms in this work. Five levels of severity for stress, depression, and anxiety were evaluated via the online DASS21 questionnaire. The machine learning techniques KNN, Naive Bayes, SVM, and Decision Tree were applied. They are particularly well suited to forecasting psychological disorders because of their accuracy. Using the various methods, it was found that classes were unbalanced in the confusion matrix. The Random Forest classifier was selected as the most accurate model of the five algorithms using the F1 score criteria. The specificity parameter also showed that the algorithms were susceptible to undesirable results. Compared to the DASS42 dataset, the dataset used in this study paper is somewhat small.

The proposed study's questionnaire-based dataset depicts the distribution of anxiety and depressive symptoms. (Singh and Kumar; 2021) used the DASS-21 questionnaire to classify anxiety and depression using machine learning algorithms on user responses. They looked through the dataset to find users who fit the depression and anxiety categories. The outputs of five classification algorithms: SVM, Decision Trees, Random Forest, Naive Bayes, and KNN, were then contrasted. SVM outperforms all other machine learning techniques. Twenty-eight variables in the dataset are solely connected to depression and stress, which is problematic because the stress was not considered in the study.

In this work, (Singh and Kumar; 2021), eight machine learning algorithms were trained to predict the emergence of psychological problems like anxiety, depression, and stress using data from the online DASS42 application. Eight algorithms were applied to forecast stress, depression, and anxiety levels. Algorithms include probabilistic, neural network, tree-based, and closest neighbour techniques. All techniques were applied to two independent databases, DASS42 and DASS21, obtained from various sources. After utilizing every strategy, the data revealed that neural networks performed better than any. The RBFN outperformed other neural networks in terms of depression in both datasets. However, the random forest result for DASS21 anxiety is 100

(Rao and Ramesh; 2015), conducted pilot research to examine the DAS level of 90 workers in a Bangalore, India-based firm to see how it affected productivity. A cross-sectional design was employed, and it was discovered that none of the workers had depression and that 18% suffered from stress and 36% from anxiety.

(S. Iqbal and Venkatarao; 2015), employed the DASS 42 questionnaire to determine whether DAS was present in medical students. She discovered that 51.3% were depressed, 66.9% were anxious, and 53% were stressed out. Additionally, they found that female students were more influenced among all undergraduates than male students, and fifth-semester students were more affected than second-semester students. They asked for early medical counselling to reduce this morbidity.

2.2 Review of previous research based on Psychological Domain and Machine Learning

The Medical College and Hospital of Kolkata, India, provided information on 630 elderly patients, with 520 receiving extraordinary care. After applying several classification methods, such as the Bayesian Network, logistic, multiple layer perceptron, Naive Bayes, random forest, random tree, J48, sequential, random optimisation, random sub-space, and K star, (Sau and Bhakta; 2017), discovered that random forest produced the best accuracy rates of 91 and 89%, respectively, among the two data sets of 110 and 520 individuals. This study emphasises the use of machine learning technology in automated screening for mental health issues. With this technology, screening procedures for severe anxiety and depression can be replaced by a computer-based method with a respectable degree of accuracy.

(Joseph et al.; 2021), goal in writing this paper is to make predictions about the emotional assessment of workers at the organization and the level of sinking personnel. The authors conducted a survey and used it to collect the data for analysis. This model predicts a decline and offers a stress analysis using the abovementioned data. The authors used Six machine learning algorithms like Decision Tree Classifier, Support Vector Machine, Random Forest, Naive Bayes, Nearest Neighbours, and logistic regression. Random Forest Classifier achieved 86% accuracy in forecasting mortality rates in this database. The authors contend that by using this model, employers and labour experts can keep track of their employees' mental health and take the necessary action to lessen weariness. This method could be used in various fields, such as banking, education, and information technology, depending on the breadth of the future. Other mental health issues, such as anxiety and depression, can be incorporated into the emotional analysis, and the model can be tailored to match the demands of different industries.

In this research, (Ahmed et al.; 2020), proposes a system that would use machine learning techniques to aid in diagnosing depression and anxiety. Using a standard scale

questionnaire, the authors polled 35,000 people aged between 18 and 35. On the pre-processed datasets for depression and anxiety, they applied CNN, K-Nearest Neighbours, Linear regression, SVM, and Linear Discriminant analysis. CNN performed well in both situations, with 96.8% and 96%, respectively, and KNN fared the worst of the five algorithms. They will create a system in the future to assist psychologists in keeping track of patients' post-therapy progress. This paper attempted an innovative strategy but was unable to determine the extent or severity of the condition or analyze other disorders.

(Sau and Bhakta; 2019), use ML technology to manage depression and anxiety in seafarers . Mental illnesses like anxiety and depression can affect people of all ages, from children to the elderly, including both men and women. The authors employ ML to identify sadness and anxiety in senior citizens or seafarers. At the Haldia Dock Complex (HDC) in India, more than 400 seafarers were questioned, and 470 persons provided the data. Multiple machine learning classifiers, such as CatBoost, Random Forest, Support Vector Machine, Logistic Regression, and Naive Bayes, may identify depression and anxiety. Catboost provides the best performance among the investigated classifiers, with maximum accuracy and precision of 82.6% and 84.1%, respectively. Fourteen features were combined into a final data set from January to July 2016. On the training set, 10-fold cross-validation is also applied.

(M. J. Patel and Aizenstein; 2015), highlighted the application of ML approaches, incorporating clinical imaging characteristics and predicting depression in older persons. The objective is to assess the precise depression model and the appropriate care and medication supplied to patients, utilizing ML techniques with model inputs (multi-modal imaging,non-imaging of the human brain). Using medicated post-recruitment, 33 patients and 35 older non-depressed individuals were each recruited separately. Their brain features and demographic and cognitive scores were obtained using multi-modal magnetic resonance imaging pretreatment. The authors show a maximum accuracy of 87.27% and a treatment response of 89.47% when testing the prediction models for depression in older persons using the alternating decision tree (ADT) machine learning method.

According to this study, (Kamite and Kamble; 2020), social media platforms can help with the early diagnosis of depression. They looked at syntactical components of Twitter tweets. They are trying to develop a system to monitor and evaluate syntactical indicators associated with the onset and maintenance of depression symptoms. They used the Random Forest technique and had a 99.89% accuracy rate. In the future, researchers intend to use deep learning models and look at more characteristics to detect not only depression but also detect any early-stage mental health issues. In this study, they didn't look at multiple variables or try various machine learning techniques. They primarily focused on the depression, disregarding other sentiments expressed in the tweets.

(Young; n.d.), identified anxiety and mood disorders by scanning patients' facial expressions and using cross-validation. To confirm that, more accurate results using several statistical metrics were discovered.

2.3 Conclusion

It is apparent from the comprehensive assessment of the literature that each of these implementations uses a different set of methodologies and datasets. All of the researchers that used the DAS dataset in this study took into account only the features of the questionnaire and ignored the general factors. Overall, it is apparent from the literature that there is excellent room for improvements and expansions in this field.

3 Research methodology

In the presented study of Depression, Anxiety, and Stress Prediction, the entire data mining process is broken down into a series of steps based on KDD methodology. Figure:1, presents each of these processes in order.

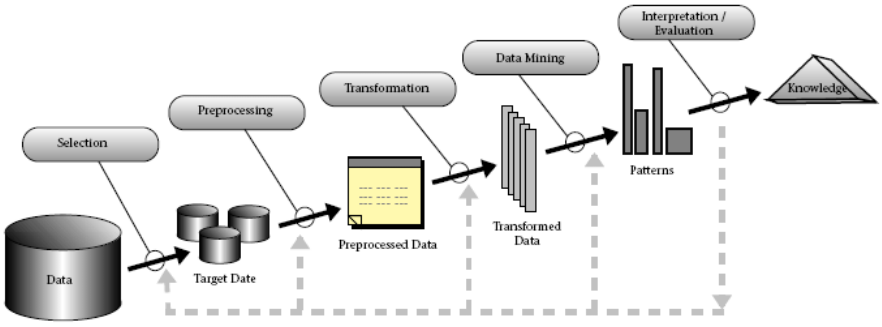

Figure 1: DAS Prediction Methodology

3.1 Dataset collection

The open-psychometrics data repository is explored in the research provided, and the data needed for the study is obtained through DASS[1], as will be discussed below.

- DASS Dataset.

  For the creation of ML models, data collection is necessary. There aren't many data collection techniques that are only based on research. Data collection methods include observation, survey, paper scanning, measurement, questionnaire, or a combination of these. Our study uses data from an online poll created by the author (Lovibond Lovibond, 1995). Through online surveys, they collected 39775 samples between 2017 and 2019.

There are a total of 42 questions in DASS42. There are no right or wrong responses in it. 14 of the 42 items fall under the category of anxiety, which assesses signs and symptoms such as persistent worry, trouble concentrating, and situations and occurrences like threatening and fatigue. Depression is 14 questions that assess symptoms, including underestimation of life, self-criticism, loss of interest, sadness, grief, inactivity, and tearfulness. The following 14 questions are divided into stress categories, which assess symptoms like trouble maintaining calm, difficulty falling asleep, sweating, overreacting, and tolerating disruptions. Figure:2, shows the text- or number-based options for each question's potential responses.

The results of an online survey were the data. It was gathered within two years and came from participants who agreed to take the research survey and indicated their consent to having their responses utilized for research by responding in the affirmative.

[1]Data Source: `http://openpsychometrics.org/_rawdata/`

scale	Meaning
0	Did not apply to me at all
1	Applied to me to some degree, or some of the time
2	Applied to me to a considerable degree, or a good part of the time
3	Applied to me very much, or most of the time

Figure 2: Rating-Scale

The dataset includes 39,775 records and 172 features, including 42 columns for survey questions and 84 factors linked to the length of time it took to answer each question and the position of the question in the survey. Along with 16 columns for the word checklist and three each for introelapse, testelapse, and surveyelapse, ten personality questions were included. There are 17 more columns for characteristics, including race, education, orientation, nation, age, gender, religion etc. The three mental diseases are compared in this dataset, which is the primary criterion for selection.

3.2 Pre-processing and transformation of data.

After the datasets are acquired, a number of pre-processing operations are carried out. Python is used to develop processing steps and model applications.

3.2.1 Dataset Cleaning

The online repository provides data in CSV format. After reviewing the data, it appears that some numbers are missing, and some inputs are not acceptable. Because the missing values are insignificant compared to the dataset size, I eliminated the unnecessary punctuation from the columns and deleted the missing rows. We should exercise caution when changing data from its original value, as it could be required to build an accurate model. Missing values may be detrimental since some algorithms will not function properly when given null values. Since the'major' column contains many null values, I gave 'No Degree' to those columns. Due to the several incorrectly spelt professions in this column, the values in the "major" column were later cleaned up by altering the data. Later converted some Object datatypes to Integer for EDA.

3.2.2 Feature Selection

I choose the most consistent, non-redundant, and pertinent features to incorporate into the model. More than half of the 172 features are not helpful for our forecast. I omitted the validity checklist and time-lapse columns and renamed the ten personality questions with personality domain names. Features showing the time spent to complete the survey and the questions positions are also removed.

3.2.3 Feature engineering and DAS dataset creation

When we leverage current features to build new features, we use feature engineering to see if those new signals can help us predict our outcome. I derived an age-group column for the age field for the EDA, which I later transformed into a numerical format for the machine learning model. The 42 questions were then isolated to a new dataset. By

dividing questions about depression, anxiety, and stress from the newly formed one, I could extract three datasets for depression, anxiety, and stress. The question numbers for depression, anxiety, and stress are respectively [3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42], [2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41], and [1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39]. I then combined the remaining attributes that were shared between each dataset, resulting in the creation of separate datasets for depression, anxiety, and stress. We generated a "Total-count" column for each dataset, representing the row-wise sum of stress, anxiety, and depression score. Next, we developed a new " condition " feature in accordance with the total score's following requirements.Figure:3 shows the DASS dataset scoring scale and Figure:4 shows the Depression score calculation.

	Depression (D)	Anxiety (A)	Stress (S)
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34+

Figure 3: DASS-42 scoring pattern.

Q3	Q5	Q10	Q13	Q16	Q17	Q21	Q24	Q26	Q31	Q34	Q37	Q38	Q42	Total score	Severity
3	3	3	3	1	2	1	3	0	3	3	3	2	3	33	Extremely severe
1	3	3	0	0	0	1	3	1	3	0	1	3	3	22	Severe
2	1	2	0	0	2	0	0	2	2	2	1	1	2	17	Moderate
0	1	2	1	0	2	1	1	0	0	0	3	1	0	12	Mild
0	0	2	0	1	0	0	3	0	1	0	3	0	0	9	Normal

Figure 4: Depression-score Calculation.

After creating the new feature 'condition', I dropped the country and age columns, which did not help to build the model. Later, a new feature named 'severity' from 'total-count' was created before dropping it. I created a severity column for making the outcome binary, which gives '1' when the condition is 'moderate' or 'severe' or 'extremely severe' otherwise ', 0' for normal and mild. Individuals with moderate mental disorders must be considered along with severe and highly severe because moderate can step towards these high levels.

3.2.4 Exploratory Analysis

Some of the most important conclusions concerning the data are presented in this section.

- Plot for top ten Majors of people and top ten countries from where people attended the survey.

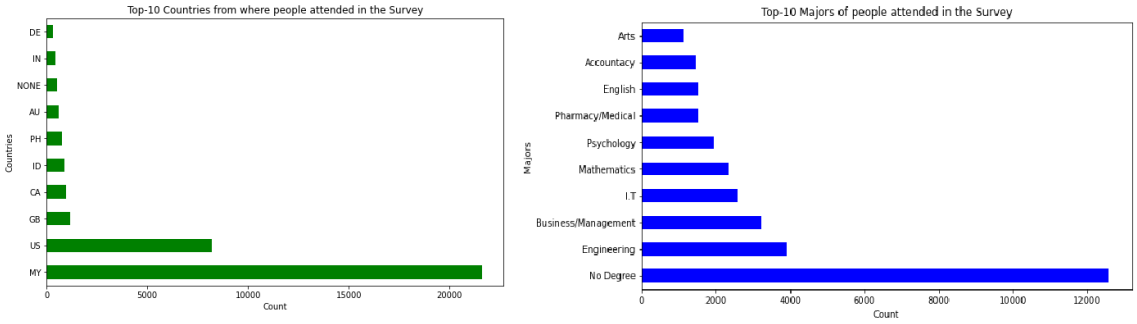

Figure 5: Top ten majors and countries

From the plot,Figure:5, we can understand that over 20K people attended the survey from Malaysia compared to nearly 9k from the United States. Most people didn't mind filling the 'majors' column, and among filled, people from engineering and Business were more compared to other 'majors'.

- Participant's conditions for Depression, Anxiety and Stress.

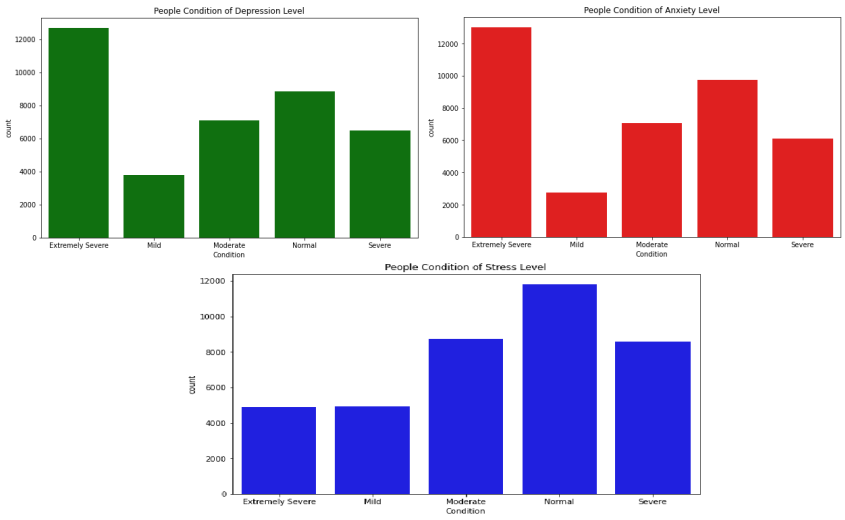

Figure 6: Participant's conditions for DAS

From the plot,Figure:6, it is understood that Most people with Depression and Anxiety are at an extremely severe level, while the Normal level is high for Stress.

- Depression- EDA of Gender,Marriage, Religion and Education Level.

For Gender (1,2,3 represents male, female and others, respectively), Over 76% of participants are female. Most people are highly depressed, but the difference between extremely depressed and normal is high in females than in males. For marriage (1,2,3 represents single, married, and divorced respectively), the Depression level is higher for singles than for married; most married people are normal compared to highly numbered depressed people among singles. For religion, from figure 6, it is understood that most

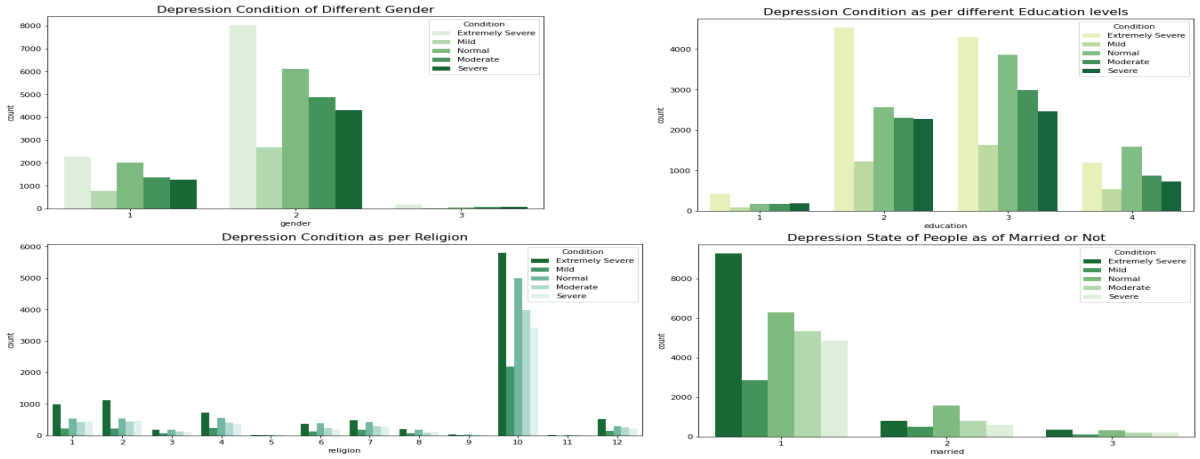

Figure 7: Depression Analysis

participants are from Malaysia, and 60% of Malaysians are Muslims, so here in this survey, Muslims are higher in number. Of all religion, extremely severe depression are more in number except for Christian(protestant), where the count of normal people is higher than all other levels. For education(1,2,3,4 represents the primary school, high school, university and graduate degree respectively); for Education level, extremely severe depression is more in number except for Graduates, where the count of normal people is higher than all other levels.Figure:7,shows the Depression Analysis plots.

- Stress- EDA of Age-groups, Races, Sexual Orientation and Religion.

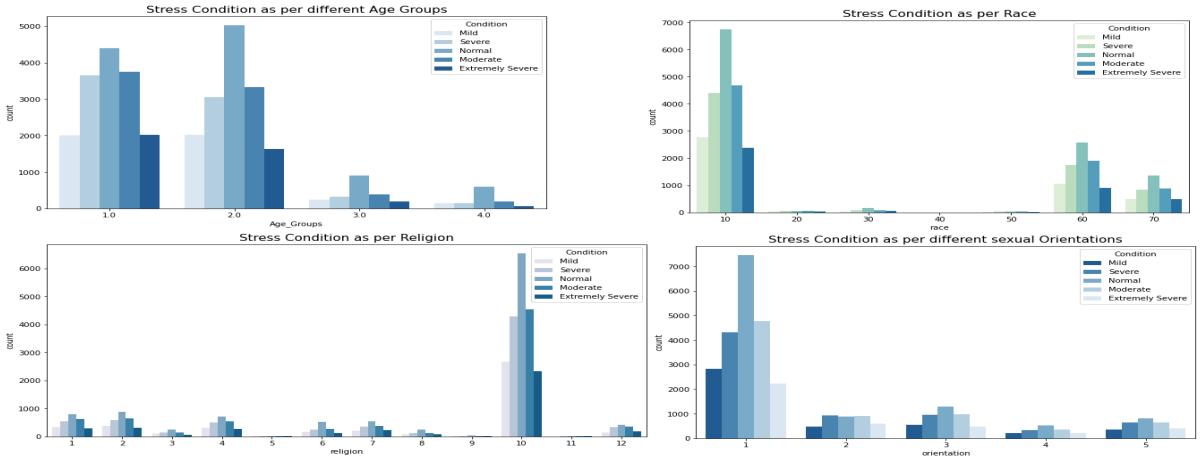

Figure 8: Stress Analysis

For Age Groups (1,2,3,4, represents age groups of (17-21), (21-35), (36-48), and (49+) respectively), stress is normal among Adults and Secondary people. Stress illness among Elder adults and Older people is also normal. For Race (10,20,30,40,50,60,70 represents Asian, Arab, Black, Indigenous Australian, Native American, White, and others, respectively). Most participants are from Malaysia; the number of Asians is higher in this survey, and most are in normal stress conditions. For sexual orientation (1,2,3,4

represents Heterosexual, Bisexual, Homosexual, and Asexual, respectively), the Stress level is normal among all sexual orientations. Among heterosexuals and the homosexual difference between normal and extremely severe stress is very large. In all religion, normal stress are more in number, where the count of normal people is higher than all other levels.Figure:8,shows the Stress Analysis plots.

- Anxiety- EDA of Education level, Marriage, Gender, Age-Group.

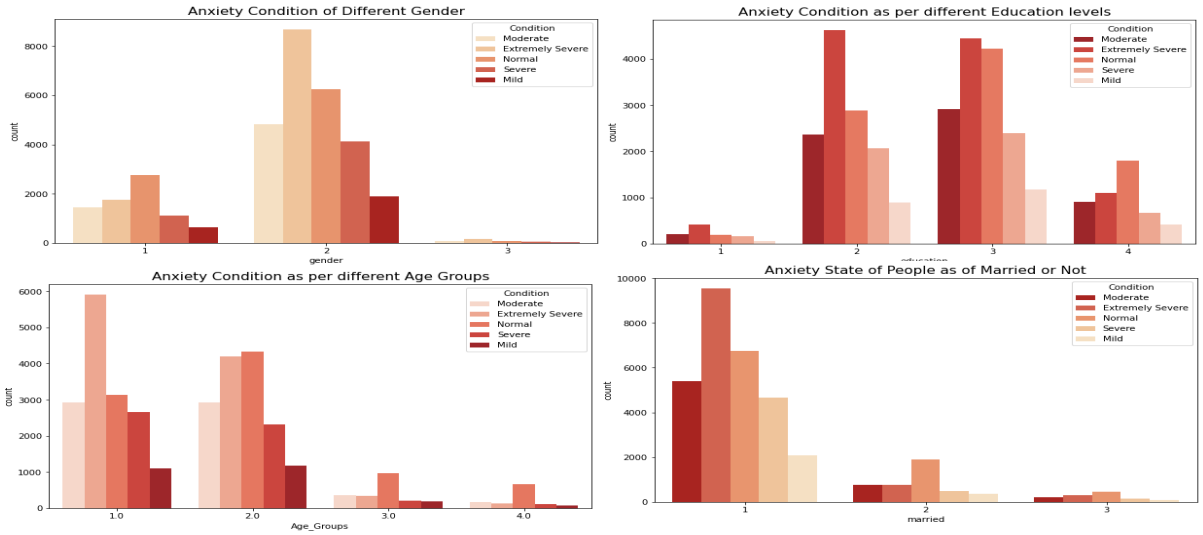

Figure 9: Anxiety Analysis

The ratio of severe people is higher for all education levels, marriage levels, Gender and all age groups. The overall proportion of anxiety among participants is at a extreme level.Figure:9,shows the Anxiety Analysis plots.

3.3 MinMaxScaler

MinMaxScaler divides the feature's range by its minimal value before subtracting it. The difference between the original maximum and original minimum called the range. MinMaxScaler keeps the original distribution's shape. The information present in the original data is not materially altered.The scale diversity of the input features frequently has an impact on how well machine learning models function. The proposed study, which changes all the features in a scale from 0 to 1, uses min-max scaling to close this gap. This is accomplished using the MinMaxScaler module from Sklearn [2], which employs the following formula:

$$xnew = ((x - xmin) = (xmax - xmin)) * (max - min) + min$$

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

3.4 Principal Component Analysis

An unsupervised, non-parametric statistical method called principal component analysis is most often used in machine learning to reduce dimensionality. The term "high dimensionality" refers to the dataset's high feature density. Model overfitting, which limits generalization beyond the examples in the training set, is the main issue connected to high-dimensionality in the machine learning area. The cumulative explained variance ratio as a function of the number of components identifies the ideal number of significant components. We plotted a chart with the number of components and cumulative variance(%) with a 95% cut-off threshold. We found that(fig:10) for the Depression dataset number of components selected is 31, but for Anxiety and Stress, it is 32.

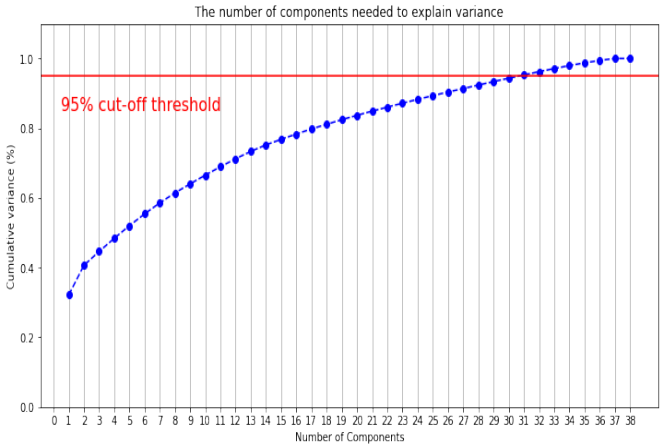

Figure 10: PCA for Depression Data

In this instance, I require 31 principle components to explain 95% of the variance.

4 Design Specification

The DAS prediction system's overall architecture, which is given in this work, Figure: 11, is separated into three primary phases, as follows:

- Database Layer.

  The DAS dataset is first downloaded in CSV format from an online data source; then, it is stored on a computer before being exported to Python. For the required analysis, pre-processing, transformation, and feature engineering, I used the Jupyter notebook. Python was chosen because it offers a variety of libraries for machine learning and data manipulation. Finally, we extracted the three necessary stress, anxiety, and depression datasets.

- Application Layer.

  DAS prediction is then performed using the processed data. Before applying algorithms, we scaled the datasets using MinMaxScaler and then applied PCA to the scaled trained and test data. The prediction is made using machine learning models like RF, Gaussian NB, and neural networks. Python-Jupyter Notebook is used to model and optimize the models that are presented.

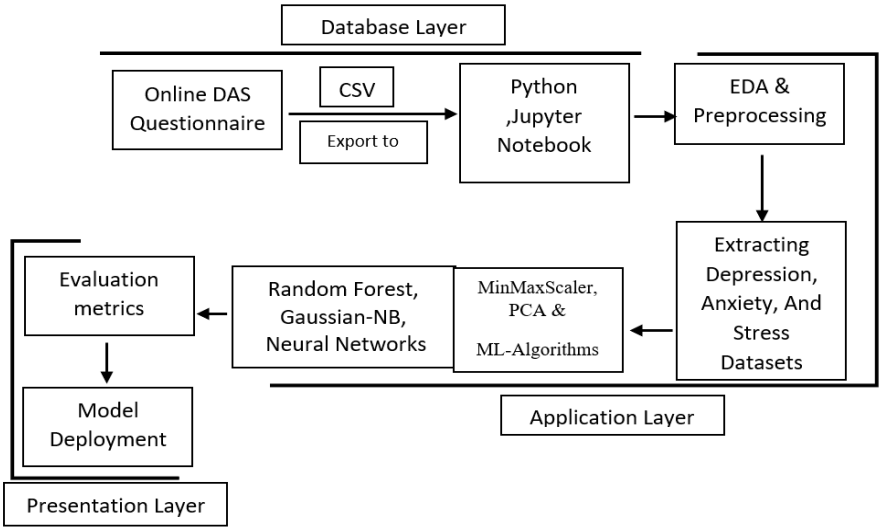

Figure 11: Design Specification

- Presentation Layer.

  Metrics that choose the optimal model are used to evaluate the outcomes. The Seaborn package included with Python is used to show the model's results.

5 Implementation

After processing and normalising the DAS dataset, I extracted the datasets for depression, anxiety, and stress from the primary dataset. The processed data has about 39775 rows prepared for the model application. The amount of input the model can learn determines how well it can be trained. As a result, the dataset is divided into 25% test and 75% train. We scaled the datasets before implementing PCA to the train and test data. To prevent bias, datasets employed random sampling. Following this, the train and test datasets are used to apply the machine learning models discussed in the following section. A careful analysis of the literature reveals that no studies consider variables like ethnicity, education, gender, sexual orientation, marriage, etc. Additionally, none of them used this dataset with neural networks. This study compares the performance of the Gaussian NB, Random Forest, and neural network models.

5.1 Machine Learning Algorithms

5.1.1 Random Forest

The bagging principle is used by the machine learning method Random Forest. It is an ensemble tree-based model that combines the performance of various subpar tree-based models to effectively understand data.This study uses the "RandomForestClassifier" package from Sklearn [3]. To achieve the intended outcomes in the case of RF, a few crucial factors must be tuned correctly. In the study that is being presented, the

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

parameters "max depth = 160" and "max features" are set to "auto," automatically set them to the number of input variables. The parameters for "n estimators," "min samples leaf," and "min samples split," in contrast, are set to 190, 1, and 3 correspondingly.

5.1.2 Gaussian Naive Bayes

The Bayes theorem is the foundation for a group of supervised machine learning classification techniques known as Naive Bayes. It is a simple categorization system, yet it works incredibly well. They are helpful when the inputs have a high degree of dimension. Complex classification problems can also be resolved with the Naive Bayes Classifier. Gaussian Naive Bayes is a variant of Naive Bayes that adheres to the Gaussian normal distribution and supports continuous data. This study uses the "GaussianNB" package from Sklearn [4]. Figure:12, Shows Gaussian normal distribution.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Figure 12: Gaussian normal distribution

5.1.3 Neural Networks

An input layer, one or more hidden layers, and an output layer are the node layers that make up neural networks in this study. Each node, or artificial neuron, is interconnected with others and comes with a weight and threshold. Any node whose output rises above the specified threshold value is activated and starts sending information to the top layer of the network. Otherwise, no data is sent to the next tier of the network. Training data is necessary for neural networks to grow and improve their accuracy over time.

The Keras library[5], which is included with the Tensorflow library, will be used in this study to create neural networks. We built a three-layer neural network with a 38-bit input structure for this study because there are 38 features to consider. A binary classification network and a multiple classification network were both constructed. Two layers, each with 16 nodes, and one output node comprise the binary classification. The final node employs the sigmoid activation function, which condenses all values between 0 and 1 into a sigmoid curve. ReLU is used as the activation function in the other two levels. ReLU is a half-rectified function; it returns 0 for any input that is less than 0 and keeps its value for positive inputs. We used Adam, a momentum-based optimizer, for model compilation. Metric used is accuracy, while 'binary crossentropy' is the used loss function. With a batch size of 32, the model is trained for 100 epochs. There are two layers, each with 16 nodes and five output nodes for multiclass prediction. While the other two levels use ReLU as the activation function, the final node uses softmax activation. We utilize the Adam optimizer to compile the models, with 'sparse categorical crossentropy' as the loss function and 'sparse categorical accuracy' as the metric. With a batch size of 32, the

[4] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

[5] https://keras.io/about/

model is trained for 100 epochs. The accuracy of the trained model was then assessed for the test set.

6 Evaluation

This section provides a critical assessment of the machine learning models' performances as they relate to the DAS prediction. The study's evaluation metrics are described below:

- Accuracy: Accuracy is the percentage of correctly predicted data points among all the data points.
- Precision: It is calculated by dividing the total number of positive predictions by the proportion of true positives.
- Recall: The percentage of positive patterns that are correctly categorized is measured using the metric recall.
- F1 score: It only quantifies the proportion of accurate predictions a machine learning model has produced.
- Cross Validation: It is the process of testing the accuracy of a machine learning model with new data.

We used Random Forest, Gaussian NB, and neural networks for each DAS dataset to develop six models. We took into account two scenarios for each algorithm, the first of which had five possible output conditions: mild, normal, moderate, severe, and extremely severe. For instance 2, we changed our dependent variables to a binary result that gives a "1" when the condition is "moderate," "severe," or "very severe," and a "0" when it is "normal," "mild," or "none."

6.1 Random Forest

Two separate cases are used to make the DAS prediction utilizing RF. In the first experiment, the RF produces output with five classes, whereas in the second experiment, it produces a binary result. The figure displays the results of this study. The performance of RF with binary outcomes appears to have improved in this case. RF performed low for anxiety when compared to stress and depression.Figure:13: Random forest result.

RF	Case	Accuracy	Precision	Recall	F1 score	Cross validation
Depression	1	92.7	92.7	92.7	92.5	92.8
	2	98	98	98	98	97.9
Stress	1	91.3	91.3	91.3	91.2	91
	2	97.3	97.3	97.3	97.3	97
Anxiety	1	89.5	89.5	89.5	89.3	89.3
	2	96.8	96.8	96.8	96.8	96.4

Figure 13: Random Forest Result

6.2 Gaussian NB

Two trials are run with GaussianNB, just like the RF. It can show that the Binary result scenario marginally increased the proposed model's accuracy. For case 2, the GaussianNB performed about the same as RF, but averagely for case 1 .Figure:14: GNB Result.

GNB	Case	Accuracy	Precision	Recall	F1 score	Cross validation
Depression	1	87.1	89.1	87.1	87.6	87.3
	2	98.1	98.1	98.1	98.1	98.1
Stress	1	85.8	87	85.8	86.1	85.7
	2	95.5	95.5	95.8	95.5	95.5
Anxiety	1	84.3	86.2	84.3	84.8	84.3
	2	95.6	95.6	95.6	95.6	95.5

Figure 14: Gaussian Naive Bayes Result

6.3 Neural networks

Similar to RF and GaussianNB, the neural network model(fig:15) has experimented with two different cases. It is observed that the neural networks model with three layers shows higher accuracy for both cases, and it performed better than RF. The accuracy of the neural network almost reached 100%. In this research, we used the accuracy metric to

NN	Case	Accuracy
Depression	1	99.7
	2	99.5
Stress	1	99.3
	2	99.5
Anxiety	1	98.7
	2	99.5

Figure 15: Neural Networks Result

compare the performance of the models for each Depression, anxiety and Stress. Figure 16 compares all models' accuracy in measuring depression, anxiety and stress.

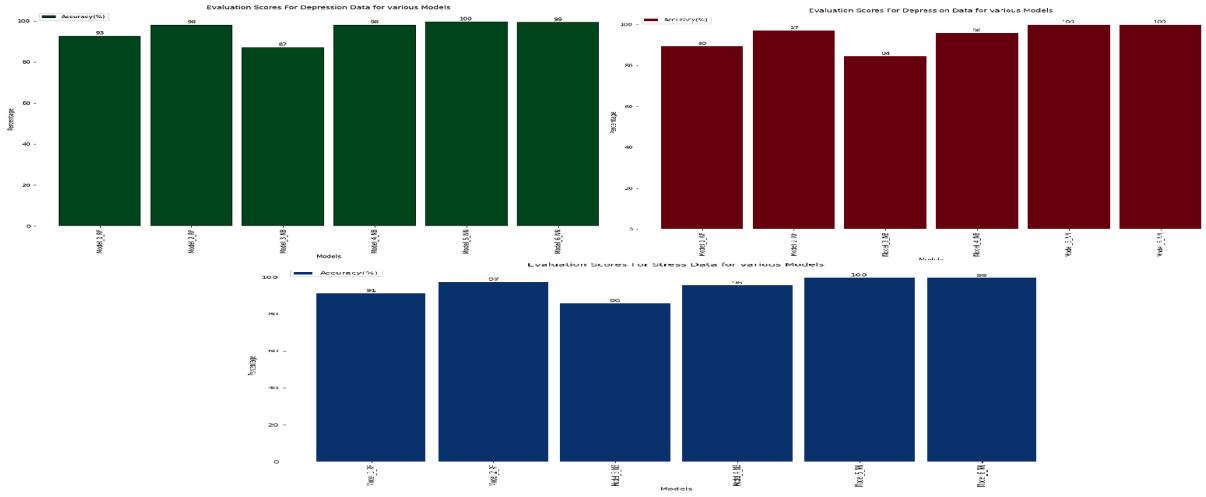

Figure 16: Accuracy comparison of models for DAS

6.4 Discussion

An overview of the results obtained is discussed in this section. The implementation of RF, GaussianNB, and Neural network models on the DASS42 dataset are described. Analyzing and contrasting the performance with that of other algorithms is essential. Because it depends on numerous variables, including the dataset's structure, size, and intended purpose, not every ML technique will be the greatest fit for every use case. Additionally, each model has a unique set of performance characteristics. Over nine ML methods, (Srinath et al.; 2022) worked. They classified depression, anxiety, and stress for various levels of severity for the dataset made up of 25% test data and 75% training data using the free and open-source "WEKA" application. They concluded that the RBFN algorithm provided the highest accuracy of all algorithms, with DAS accuracy rates of 96.02%, 97.48%, and 96.17%. (Priya et al.; 2020) worked on five Ml algorithms, and they discovered that Naive Bayes provided them with the highest accuracies of 85.5%, 73.3%, and 74.2% for DAS. The researchers cited above didn't consider general variables like religion, education, health, gender, etc. On the same dataset with the same proportion used for training and testing, we built RF, GaussianNB, and Neural Networks in Python while considering 14 general parameters and ten personality traits. We converted 'normal' and 'mild' as 0 and 'moderate', 'severe', and 'extremely severe' as 1 for case 2; by doing this, we converted five output classes to a binary output. When we implement RF, Gaussian NB and Neural network on the binary classified dataset, the accuracy of RF and GaussianNB increased by nearly 5% and 10%, respectively, for DAS. In addition, we identified that 14 general parameters and ten personality traits have a major contribution to DAS prediction, improving the prediction accuracy. Overall, this study has presented valuable contributions to the field of Depression, Anxiety and Stress prediction. The accuracy of the models can help the psychology domain effectively predict mental illness. Neural networks nearly give an accuracy of 100%. This level of accuracy is significant for the healthcare industry as it demonstrates that the negative cases (people without diseases) are correctly classified. Every algorithm used in this investigation generated incredibly accurate results for negative instances.

7 Conclusion and Future Work

The main goal of this work was to predict the DAS using DASS42. We implemented RF, GaussianNB and Neural networks for our study using python. By studying the previous works, we considered 14 general parameters and ten personality traits along with questionnaire features; previous researchers did not consider these 24 additional features. Neural Networks were identified as the best model with this dataset in both cases, and Neural networks achieved an accuracy of 99.5% and 99.7% for case 1 and case 2, respectively. Random Forest and GaussianNB achieved higher accuracy for case 2 than case 1; both algorithms got their lowest accuracy for the anxiety dataset. From the exploratory analysis of the data, it is observed that most participants are from Malaysia and US. When we analyzed the remaining factors, we found that most participants are Muslims and of Asian ethnicity, which indicates a bias of the participants, and since it is an online survey, the number of adults and secondary people is large compared to old and primary participants in the survey. The stress ratio among each group for every factor is normal, but Depression is extremely severe for almost all groups but normal for married, elderly and graduated individuals. In the case of Anxiety, the ratio is normal for married

people and females, but for every other factor, it is extremely high. The unique findings from this study will also help the other scholars in their future DAS research endeavours.

In addition to psychological issues, the standard of living, the area where individuals live, and other socioeconomic aspects also play a role in mental illness. The given study for the DAS prediction system based on more intricate deep learning models taking into account more qualitative lifestyle characteristics that can identify DAS early will be intriguing to extend in future work.

8 Acknowledgement

The author thanks the help that Mr. Parmod Pathak, Mr. Paul Stynes,and Ms. Musfira Jilani provided in order to complete this work. Their knowledge aided in improving both the report-writing and technical skills.

References

Ahmed, A., Sultana, R., Ullas, M. T. R., Begom, M., Rahi, M. M. I. and Md. Ashraful Alam, P. (2020). A machine learning approach to detect depression and anxiety using supervised learning, *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering* p. 6.

Beck, A. T., S. R. A. and Garbin, M. G. (1988). "psychometric properties of the beck depression inventory: Twenty-five years of evaluation, *Clin. Psychol. Rev.* **8**.

Chen X, Li H, Z. X. H. J. (2021). "effects of music therapy on covid-19 patients' anxiety, depression, and life quality: a protocol for systematic review and meta-analysis", *Medicine* **100**(26).

Joseph, M. R., Udupa, M. S., Jangale, M. S., Kotkar, M. K. and Pawar, M. P. (2021). Employee attrition using machine learning and depression analysis, *2021 5th International Conference on Intelligent Computing and Control Systems* p. 6.

Kamite, S. R. and Kamble, D. V. (2020). Detection of depression in social media via twitter using machine learning approach, *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing* p. 4.

Kaur, G. H. Tee, S. A. and Krishnapillai, A. S. (2013). "depression, anxiety and stress symptoms among diabetics in malaysia: a cross sectional study in an urban primary care setting.

M. J. Patel, C. Andreescu, J. P. K. E. C. R. and Aizenstein, H. J. (2015). Machine learning approaches for integrating clinical and imaging features in late - life depression classification and response prediction, *International journal of geriatric psychiatry* **30**(10): 1056–1067.

Priya, A., Garg, S. and Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms, *Procedia Computer Science* **167**: 1258–1267.

Radloff, L. (1977). The ced-d scale: A self-report depression scale for research in the general population, *Applied Psychological Measurement* **1**: 385–401.

Rao, S. and Ramesh, N. (2015). Depression, anxiety and stress levels in industrial workers: A pilot study in bangalore, india, *Ind. Psychiatry journal* **24**: 23–28.

S. Iqbal, S. G. and Venkatarao, E. (2015). stress, anxiety depression among medical undergraduate students their socio-demographic correlates, p. 354–357.

Sau, A., B. I. (2017). Predicting anxiety and depression in elderly patients using machine learning technology, *Healthcare Technology Letters* p. 6.

Sau, A. and Bhakta, I. (2017). Predicting anxiety and depression in elderly patients using machine learning technology, *Health technology Letters* .

Sau, A. and Bhakta, I. (2019). Screening of anxiety and depression among the seafarers using machine learning technology, *Informatics in Medicine Unlocked* **16**: 100–149.

Singh, A. and Kumar, D. (2021). Identification of anxiety and depression using dass-21 questionnaire and machine learning, *2021 First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT)* pp. 69–74.

Srinath, K. S., Kiran, K., Pranavi, S., Amrutha, M., Shenoy, P. D. and Venugopal, K. R. (2022). Prediction of depression, anxiety and stress levels using dass-42, *2022 IEEE 7th International conference for Convergence in Technology (I2CT)* **1**: 1–6.

WFMH (n.d.). "depression: a global crisis", *World Federation for Mental Health,World Health Organization. Accessed 4 Dec 2017* .

Young, C., H. S. B. T. W. L. (n.d.). Using machine learning to characterize circuit-based subtypes in mood and anxiety disorders, *Biological Psychiatry 85* **10**.