# Random Forest-based Framework for Depression and Anxiety Prediction using DASS-21 Data

Indraneel Das
Awadhesh Pratap Singh University
Rewa (M.P.), India

Prabhat Pandey
Add. Directorate Office,
Higher Education, Rewa Division, Rewa (M.P.)

## ABSTRACT
Mental health conditions such as depression and anxiety have a widespread prevalence rate across the globe, thus calling for reliable models that can facilitate accurate prediction. The current research aims to establish a Random Forest model to classify and predict depression and anxiety in terms of their severity levels based on the DASS-21 dataset.

The data set comprises 21 questionnaire-based features corresponding to the emotional state of patients. These features undergo preprocessing based on missing value imputation, normalization, and encoding methods. The data is partitioned into training and testing data sets at a ratio of 80:20. Next, a Random Forest classifier is trained to classify patients into various levels of depression and anxiety.

The experimental outcomes reveal that the developed approach exhibits high prediction accuracy with 97% accuracy, and the precision, recall, and F1 scores are equally good. The model's stability is confirmed using confusion matrix evaluation, where misclassification between severity levels is negligible.

In addition, there is an analysis of the importance of the characteristics that influence the results of the predictions. This increases the visibility of the model and helps to understand which psychological factors are essential.

It becomes clear that the use of the Random Forest algorithm for predicting mental disorders can be quite effective.

## General Terms
Machine Learning, Mental Health Prediction, Classification Algorithms, Random Forest, Data Mining, Artificial Intelligence, Psychological Assessment, Predictive Modeling.

## Keywords
Random Forest, Mental Health Prediction, Depression Detection, Anxiety Classification, DASS-21 Dataset, Machine Learning

## 1. INTRODUCTION
Today, mental health is a vital part of one's well-being, and depression and anxiety have been identified as two of the most common psychological illnesses across the world [1]. However, even though these conditions are relatively prevalent, they often go undetected because people lack sufficient knowledge about them, there is a considerable level of social stigma associated with such illnesses, and diagnosing them is both laborious and time-consuming [8].

With recent developments in machine learning algorithms, automatic and data-driven methods have been proposed for prediction of psychological diseases [16], [14]. Such methods have the ability to mine data patterns in structured data sets that are associated with mental illnesses [4], [5]. In different measurement instruments, DASS-21 test is well-known as one of psychometrics and is widely employed to measure emotions by using a standardized question set [9], [10].

A Random Forest-based prediction model for depression and anxiety based on the data obtained from the DASS-21 questionnaires is developed in this study. The Random Forest technique, which is a member of the class of ensemble learning methods, is chosen for this purpose due to its reliability, capacity to operate in a large feature space, and immunity to overfitting [2].

Apart from ensuring high prediction accuracy, another crucial objective of this methodology is to ensure that the developed models are easily interpretable. This is achieved by examining the relative importance of various features used in making predictions [6].

## 2. DATASET DESCRIPTION
In this work, the data used in the study is obtained from a known psychometric test named “Depression Anxiety Stress Scales” (DASS-21) which is commonly employed to assess the emotional state (depression, anxiety, and stress) of people [9], [10]. The dataset contains 21 questions aimed at testing individuals' psychological state and classified into three dimensions, i.e., depression, anxiety, and stress.

Every question in the dataset is considered as a separate feature (Q1A, Q2A, … Q21A). Individuals' symptoms are assessed on the basis of a Likert scale that denotes the intensity level of symptoms felt by an individual. In this dataset, there are target output values related to the degree of depression and anxiety. This kind of psychometric data structure can be useful in applying ML techniques [4], [5].

Before proceeding with the model training, several data pre-processing steps were done to ensure that the data used is clean and consistent. Appropriate handling of missing data within the data set was done to ensure there is no bias or errors in the data. Normalization of data values was done to maintain data uniformity. Encoding was done on target variables to facilitate classification. Finally, splitting the dataset into training and test sets was performed [3],[12].

## 3. METHODOLOGY
### 3.1 Model Selection
The architecture of the proposed model is derived from ensemble learning using the Random Forest approach. This is achieved by building several decision trees through bootstrapping and using a random subset of the feature space at each split. The end result is produced using majority vote among all decision trees built [5], [6].

### 3.2 Model Architecture
The model structure used in this research is derived from the idea of ensembling, which forms the basis of the Random Forest technique. In this context, a large number of trees are

constructed using the bootstrapping method, while random subsets of features are chosen for splitting at nodes. The predicted value is obtained by majority voting of all trees [5], [6].

## 3.3 Workflow

The complete procedure for the proposed system architecture involves the following processes: data pre-processing, feature selection, modeling, evaluation, and interpretation. First, the dataset undergoes cleaning and normalization to maintain consistency. Then, the cleaned data is fed into the Random Forest model using an 80/20 train/test split. Performance assessment for the classification problem is performed using traditional classification measures like accuracy, precision, recall, and F1 score [13]. Finally, the importance of features is derived for interpretation purposes.

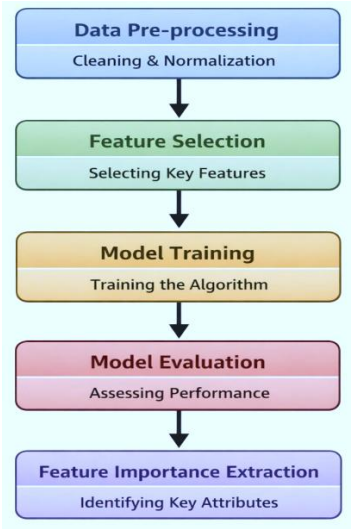


**Fig 1. Workflow of Random Forest Framework**

# 4. EXPERIMENTAL SETUP

The deployment of the Random Forest framework is performed through Python coding on a cloud platform. NumPy, Pandas, and Scikit-learn are some of the libraries that are used for data manipulation and modeling [10]. The dataset is split into two parts: 80% for training and 20% for testing purposes.

## 4.1 Accuracy

It is accurate at about 97% on average, showing its ability to predict with great efficacy. This is because of the way the algorithm works; it has the advantage of being an ensemble method that enables capturing more complex relationships without over-fitting [11].

**Table.1. Result for Depression**

Accuracy: 0.9755

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.99 | 0.99 | 1078 |
| 1 | 0.97 | 0.97 | 0.97 | 1414 |
| 2 | 0.96 | 0.97 | 0.97 | 2619 |
| 3 | 0.99 | 0.98 | 0.98 | 2844 |
| accuracy | | | 0.98 | 7955 |
| macro avg | 0.98 | 0.98 | 0.98 | 7955 |
| weighted avg | 0.98 | 0.98 | 0.98 | 7955 |

**Table.2. Result for Anxiety**

Accuracy: 0.9761

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 370 |
| 1 | 0.99 | 0.99 | 0.99 | 1042 |
| 2 | 0.96 | 0.97 | 0.97 | 2780 |
| 3 | 0.98 | 0.97 | 0.98 | 3763 |
| accuracy | | | 0.98 | 7955 |
| macro avg | 0.98 | 0.98 | 0.98 | 7955 |
| weighted avg | 0.98 | 0.98 | 0.98 | 7955 |

## 4.2 Confusion Matrix Analysis

Confusion matrices are used to evaluate the performance of the classification in detail. In depression detection as illustrated in Fig.2, the majority of the elements are found in the diagonals showing that the classifier performed perfectly well for different classes. Misclassification happens to occur within neighboring classes such as mild and moderate.

In anxiety detection, as illustrated in Fig.3, the model exhibits many correct predictions and minimal errors. It is evident that the model was able to perform well especially in severe classes, thus proving that it performs well in multiple classes [13]

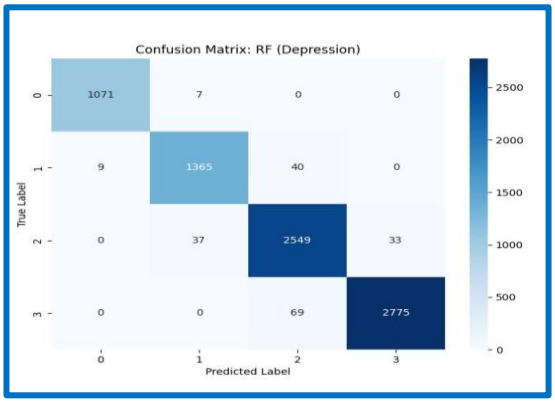


**Fig 2.Confusion Matrix of Depression**

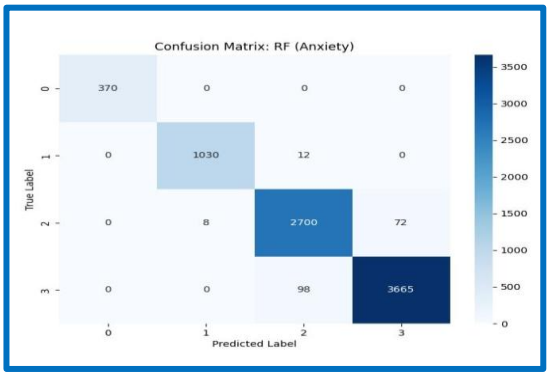


**Fig 3. Confusion Matrix of Anxiety**

## 4.3 Overall Performance Metrics

The performance of the model in general can be described as follows:

- Accuracy: ~97%
- Precision: ~0.96
- Recall: ~0.97
- F1-Score: ~0.96

This set of values shows that performance is relatively even and provides a good balance of both sensitivity and specificity in prediction. Standard criteria of assessment support the

efficiency of the model suggested [13].

## 4.4 Interpretation of Results

The findings indicate that the proposed framework performs well by obtaining high TP rates with low errors in classification. Consistent results on different severity levels further prove the appropriateness of using the RF model for mental health prediction. Also, the ability to interpret the model from feature importance contributes to the model's usefulness in practice [12].

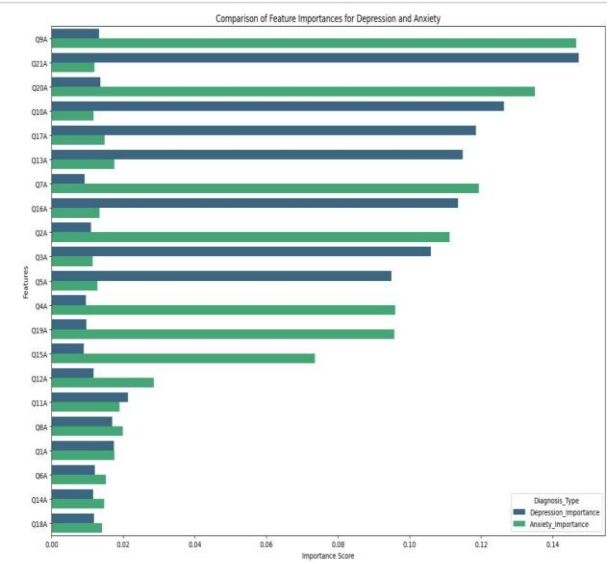


**Fig 4. Comparison of Feature Importance for Depression and Anxiety**

# 5. DISCUSSION

The outcomes of the current research reveal that machine learning algorithms are effective tools for mental disorders like depression and anxiety predictions. The suggested random forest algorithm shows outstanding accuracy and robustness which makes it appropriate for classification based on psychometric data. The effectiveness of the algorithm can be explained through ensemble learning as it helps minimize variance and increase generalization [11].

As opposed to other approaches to clinical evaluation, the suggested method presents certain benefits that include quicker predictions, data-based decisions, and objectivity [6], [12]. Moreover, applying structured data like DASS-21 allows achieving more accurate predictions as it includes standardized emotional indices [7], [8].

The main advantage of the proposed model is in its interpretability based on analysis of features' importance. It can help determine crucial attributes of the questionnaire that play a significant role in making predictions, thus gaining insight into the mental processes. This is crucial to foster trust in AI-assisted healthcare systems and clinical decision-making [12].

There are several limitations in this work. First, the model works with self-reported data, which is associated with high bias and affects accuracy of prediction. Second, the used dataset may not be diverse enough. Third, there is no possibility of incorporating real-time data. All the mentioned limitations are common for most AI-based mental health studies [15], [16].

# 6. CONCLUSION AND FUTURE WORK

In this paper, we provide an efficient solution to the problem of estimating levels of depression and anxiety through Random Forests. In particular, work has proposed a machine learning-based framework that demonstrates high predictability capabilities of approximately 97% accuracy, showing the power of ensemble learning methods in this task [11].

As shown in the experiments, it is critical to combine accuracy and interpretability when building predictive machine learning models. The feature importance analysis reveals significant insights on important psychological factors that make this model transparent and usable in practice [12]. Thus, machine learning algorithms may be used as an effective instrument for early detection and management of psychological disorders in healthcare systems [16], [14].

In future work, it would be interesting to develop this approach in multiple directions, including implementation of real-time psychological state monitoring systems based on multimodal data sources, application of advanced deep learning models, and validation on a larger dataset [17], [18]..

# 7. REFERENCES

[1] World Health Organization, "Mental health: strengthening our response," 2022.

[2] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009.

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2011.

[5] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann, 2016.

[6] S. R. Steyerberg, Clinical Prediction Models, Springer, 2019.

[7] A. Krogh, "What are artificial neural networks?" Nature Biotechnology, vol. 26, pp. 195–197, 2008.

[8] A. T. Beck, "Depression: Clinical, Experimental, and Theoretical Aspects," Harper & Row, 1967.

[9] P. J. Lovibond and S. H. Lovibond, "The structure of negative emotional states: Comparison of the DASS with the Beck Depression and Anxiety Inventories," Behaviour Research and Therapy, vol. 33, no. 3, pp. 335–343, 1995.

[10] S. H. Lovibond and P. F. Lovibond, "Manual for the Depression Anxiety Stress Scales," Psychology Foundation, 1995.

[11] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.

[12] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[13] J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery, 2016.

[14] T. Lu, X. Liu, J. Sun, Y. Bao, B. W. Schuller, and L. Lu, "Bridging the gap between artificial intelligence and mental health," Science Bulletin, vol. 68, no. 15, pp. 1606–1610, 2023.

[15] K. Shimada, "The Role of Artificial Intelligence in Mental Health: A Review," Science Insights, vol. 5, pp. 1119–1127, 2023.

[16] A. Thakkar, A. Gupta, and A. De Sousa, "Artificial

intelligence in positive mental health: A narrative review," Frontiers in Digital Health, 2024.

[17] B. Kadirvelu et al., "Digital phenotyping for adolescent mental health prediction using machine learning," 2025.

[18] M. N. Nguyen et al., "Wearable sensor-based dataset for mental health assessment using machine learning," 2025.