DOI. No. 10.13052/2794-7254.020

Predictive Analysis of Mental Health Using Machine Learning for Depression Prediction

Swati Mishra and Divyanshi Srivastava*

Abstract: This work aims to predict depression based on diverse data using Machine Learning Algorithms. The designed model seeks to identify early indicators of depression, providing a potential tool for proactive intervention and support in mental health by analyzing patterns in behavioral, physiological, and contextual data. Machine learning algorithms, namely decision trees, extra trees, XGBoost, Stochastic gradient descent, grid search CV, Stacking, and Voting classifiers, etc., are used to predict depression in the early stage.

This study emphasizes integrating machine learning techniques to enhance predictive accuracy and contribute to developing accessible and timely depression detection systems. The F1 score was added, which helped to identify the best machine learning algorithm among the ones applied. We have achieved an accuracy of 92 % with random forest, which is 3% higher than the work previously done in RF. We also achieved a 0.99 F1 score using Linear SVM.

Keywords: Machine learning algorithms, depression dataset, grid search.

1. Introduction

In contemporary society, mental health concerns, particularly depression, have emerged as significant public health challenges. The pervasive impact of depression underscores the urgency of developing innovative approaches for its early detection and intervention. In this project, our goal is to address this pressing issue by harnessing the capabilities of machine learning techniques for the early prediction of depression. By leveraging diverse data, including behavioral patterns, physiological signals, and contextual information, the goal is to discern nuanced patterns indicative of depressive symptoms.

Data collected various information on the patient, including demographics, medical conditions, history, drug use, prescription medications usage, etc. Depression is often characterized by persistent feelings of sadness and disinterest and can have profound consequences on an individual's well-being. Traditional diagnostic methods rely heavily on subjective assessments and self-reporting, leading to delays in identification and intervention. The proposed machine learning model seeks to overcome these limitations by extracting insights from comprehensive datasets, enabling the detection of subtle markers that may precede clinical manifestations. Countless individuals worldwide grapple with depression, a debilitating condition that can disrupt normal and joyful living.

From challenging daily life to the severe outcome of suicide, the impact is profound. Primarily characterized by persistent feelings of sadness and disinterest, it can have profound consequences on individuals' well-being. Traditional diagnostic methods rely heavily on subjective assessments and self-reporting, leading to delays in identification and intervention. The proposed machine learning model seeks to overcome these limitations by extracting insights from comprehensive datasets, enabling the detection of subtle markers that may precede clinical manifestations.

2. Preliminaries

Machine learning algorithms are used to predict depression using available datasets. Research has been done using different machine learning algorithms on datasets by researchers to predict depression. The outcome of this research could pave the way for proactive mental health strategies, providing individuals with the support they need before symptoms escalate. M. Keerthiga et al. [2] presented Machine Learning-based Depression Prediction using Social Media Feeds. They used a Decision tree model using a count vectorizer and achieved 89.19% accuracy and a recall of 89.85%.

The proposed approach in [3] concentrated on predicting using a Logistic Regression machine learning algorithm. The authors attained a Precision of 83% and an F1-score of 91%. N. T. Singh et al. [5] focused on stress detection from bio-signals such as heart rate variability (EEG, ECG, and HRV) and performed experiments using Machine Learning Techniques. Their results showed that the degree of accuracy depended on the size of the clinical dataset collected. A. Btabyal et al. [6] used different machine learning algorithms like DT, LR, RBF-SVC, KNN, RF, XGB, L-SVC, NB, and SV Con the scaled dataset using Standard Scaling. Out of which LR, KNN, and SVC outperformed other classifiers. D. Shi et al. [8] performed experiments on RF, DT, and SVM machine-learning algorithms. They attained an F1 score of 0.71 and an RMSE of 4.21. Mishra et al. [15] focused on cancer classification using Machine learning techniques. Mishra et al. [20] focused on the classification of histopathological cancer images using deep learning models. S. Mishra et al. [31] focused on the skin cancer classification using CNN. They achieved the highest accuracy with the MobileNetv2 model. Mishra et al. [32] paid attention to the early detection of depression using various machine-learning techniques. The methodology is discussed in Section 3.

JSS Academy of Technical Education, Noida, India
E-mail: swati.jss2008@gmail.com; dvyasi1309@gmail.com
*Corresponding Author
Manuscript received 04 April 2025, accepted 14 July 2025, and ready for publication 15 August 2025.

Different machine learning algorithms used for classification are represented in Section 4. Section 5 deals with the experimental results and discussions. Lastly, the conclusion of the work is summarized in Section 6.

3. Methodology

This project focused on predicting depression from the collected data of patients. The classification was done using different machine learning algorithms, namely, Random Forest, K Nearest Neighbor, Decision Tree, Naïve Bayes, SVM (Linear and Polynomial), Logistic Regression, Extra Trees, XGBoost, Stochastic Gradient Descent, Grid Search CV, Stacking, and Voting Classifier. Figure 1 illustrates the block diagram of the work done in this paper.

3.1. The Dataset Description

The depression dataset used in this research was obtained from the Centers for Disease Control and Prevention. Data was filled by the participants using a questionnaire and comprised a variety of information, including demographic, medical (age and cancer), physical, and history of the patient. Birthplace, veteran status, and household income are considered in demographic data. Arthritis, body measures, blood pressure, cholesterol, alcohol consumption, sleep disorders, smoking, etc., are also taken into consideration as the parameters to measure the mental health of a patient. This data is released every two years [11]. The nature of the outcome variable is a Binary Class.

3.2. Data Pre-processing

It is pivotal for improving the accuracy of the prediction models. The following pre-processing techniques were applied to the data.

- **Feature Engineering:** Feature engineering involves transforming raw data into a format that improves a machine learning model's performance. It includes selecting relevant features, creating new ones, and optimizing existing ones to enhance the model's ability to learn patterns and make accurate predictions.
- **Filling missing values:** To emulate the way all information may not be available for every patient, missing values were filled as "missing" or 0. Instances where people refused to answer are treated as null values and filled with missing or 0. The null values are filled with empty strings across the rows to a single column to add all the values into one.
- **Scaling:** Scaling in machine learning refers to the process of standardizing or normalizing the features of a dataset. Scaling helps improve the convergence of optimization algorithms and enhances the model's performance.

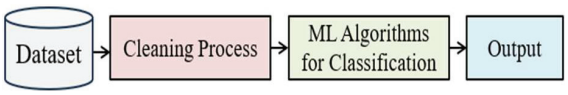

Figure 1.
Block diagram.

Table 1.

Details of the specification	
Model	**Specifications**
Used Software	Jupyter (6.5.2)
Language	Python
Kernel	Python 3.11.5 (ipykernel)
Library	Scikit-learn (version 1.3.2)
Framework	OSEMN
Number of Parameters Used	15

3.3. Model Training and Testing Strategy

The machine learning process begins with data collection and preprocessing, ensuring its quality and relevance. Next, a suitable model is selected, and a train-test split is applied. The model is trained on the depression dataset, adjusting its parameters to minimize errors and optimize performance. Different types of classifiers are used to make predictions. Finally, the trained model is deployed for making predictions or decisions in production environments. A model is trained and then fitted with various default parameters as a base. The dataset is divided into two subsets using a train-test split. The split is around 80% for training and 20% for testing. One-hot encoding is used for model preparation, especially as the dataset deals with categorical variables. One-hot encoding helps convert these categories into a numerical format that machine learning models can understand. The quantile transformer is used for scaling the data. K-means clustering is done on the data and added as a feature for modeling. For evaluating model performance, functions were written to run a classification report, make a confusion matrix, plot an ROC curve, and plot feature importance in the case of tree-based models. The specification chosen to perform experiments is shown in Table 1.

3.4. Proposed Model

Our proposed model is implemented in the following steps, as shown in Figure 2.

Step 1: The dataset is taken from the Centers for Disease Control and Prevention [11].

Step 2: Data Preprocessing begins by preparing the dataset, handling missing values, encoding categorical variables, and scaling features to ensure data quality and uniformity.

Step 3: Choose appropriate classifiers based on the problem's nature and data characteristics. Common choices include K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, Decision Trees, and Support Vector Machines (SVM).

Step 4: The performance of the model is assessed using various evaluation metrics.

4. Classification Using ML Algorithms

Different classifiers, namely RF, KNN, NB, SVM, DT, ET, LR, XGB, SGD, LR Grid Search, Stacking, and Voting classifiers, were implemented for the classification of the dataset.

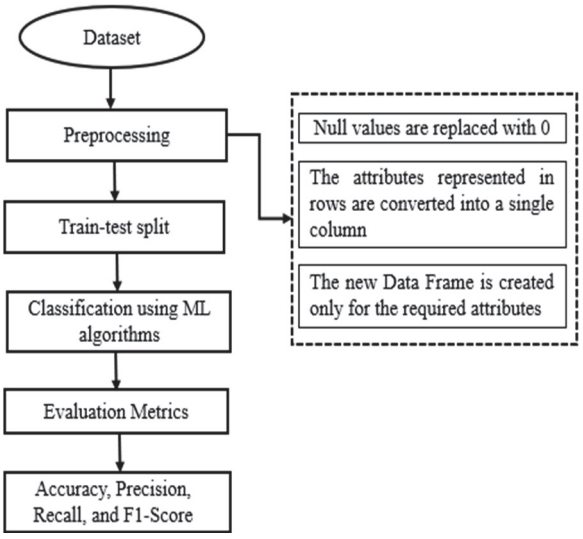

Figure 2.
Flow of implementation of our work.

4.1. Random Forest

RF creates multiple decision trees during training and combines their predictions for classification. Individual trees are trained on a random subset of the dataset. This reduces the overfitting of the data and increases accuracy.

4.2. K Nearest Neighbors

KNN assesses the labels of K-nearest neighbors in the training set. Widely used in supervised learning for pattern recognition, data mining, and intrusion detection. Choosing an odd K value helps avoid ties in classification, and cross-validation aids in determining the optimal K. Distance metrics like Euclidean, Manhattan, and Minkowski are employed to identify closest neighbors for query points that are written in Equations (1), (2), and (3).

$$\text{Euclidean distance } (x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_i)^2} \tag{1}$$

$$\text{Manhattan distance: } d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

$$\text{Minkowski Distance: } d(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p \right)^{\frac{1}{p}} \tag{3}$$

4.3. Support Vector Machine

The Support Vector Machine (SVM) identifies an optimal hyperplane in an N-dimensional space using the data points of different attributes. The hyperplane then maximizes the closest points of different attributes. Linear and Polynomial SVM were implemented in this project.

4.4. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) uses stochastic gradient descent optimization. SGD finds an optimal decision boundary by separating data points into different classes in a feature space.

4.5. Extreme Gradient Boosting

XGBoost uses gradient-boosted decision trees for classification. It works by combining predictions of individual trees and sequentially adding weak learners whilst correcting errors made by existing ones.

5. Results and Discussion

To measure the performance of the developed model, we consider accuracy (ACC), precision (Pre), recall (Rec), and F1-score metrics computed along with the confusion matrix, as shown in Figure 2. The following Equations (4), (5), (6), and (7) represent the formulations of the metrics. The confusion matrices and ROC curves of each algorithm was also plotted for evaluation of the performance of machine learning algorithms, as shown in Figures 2, 3, 4, and 5, respectively.

$$\text{Accuracy} = \frac{\text{Sum of diagonals (TP)}}{\text{Total number of instances}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Confusion Matrix – Figure 3 shows the confusion matrix drawn between actual and predicted classes. In this TP, FN, FP, and TN tell us about the True Positive, False Positive, and True Negative.

We obtained the following results for machine learning algorithms after performing our experiments on the depression dataset, as shown in Tables 2, 3, 4, and 5 for accuracy, precision, recall, and F1 Score, respectively. RF, KNN, SVM, XGB, SGD, Stacking, and Voting classifiers were applied, and their accuracies were compared. RF achieved an accuracy of 92% as compared to [12]. Polynomial SVM achieved an accuracy of 85% as compared to [14]. KNN

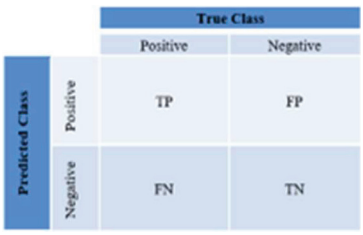

Figure 3.
Confusion matrix.

Table 2.

Accuracy comparison		
Classifiers	**References Accuracy %**	**Proposed Accuracy %**
Random Forest [12]	89	92
KNN [13]	84	86
Linear SVM [14]	85	92
Polynomial SVM [14]	85	85
XGBoost [12]	64	87
SGD	–	79

Table 3.

Precision comparison		
Classifiers	**References Precision %**	**Proposed Precision %**
Random Forest	–	92
KNN [13]	77	91
Linear SVM [14]	89	93
Polynomial SVM [14]	89	95
Decision Tree	–	91
XGBoost	–	94
SGD	–	96

Table 4.

Recall comparison		
Classifiers	**References Recall %**	**Proposed Recall %**
Random Forest	–	100
KNN	–	92
Linear SVM [14]	85	99
Polynomial SVM [14]	85	89
XGBoost	–	91
SGD	–	82

Table 5.

F1 Score comparison		
Classifiers	**References F1 Score %**	**Proposed F1 Score %**
Random Forest [12]	80	96
KNN [13]	77	92
Linear SVM [14]	85	96
Polynomial SVM [14]	85	92
XGBoost	–	93
SGD	–	88

achieved an accuracy of 86% as compared to [13]. We have also performed experiments using SGD. We achieved an accuracy of 79 % for SGD. We attained good precision in comparison to references [13,14].

Table 3 shows that we achieved the highest precision of 0.96 with SGD. We achieved the highest recall value of 0.99 with linear

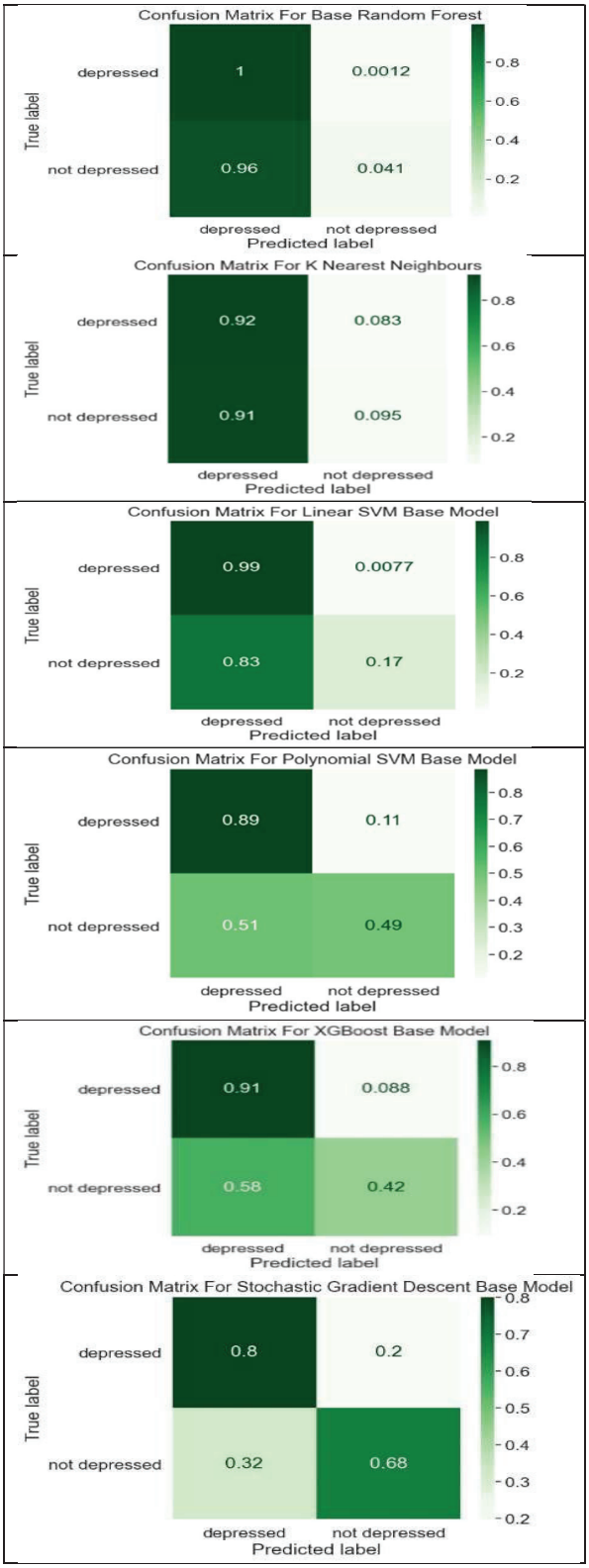

Figure 4.
Confusion matrix of RF, KNN, SVM, XGB, and SGD.

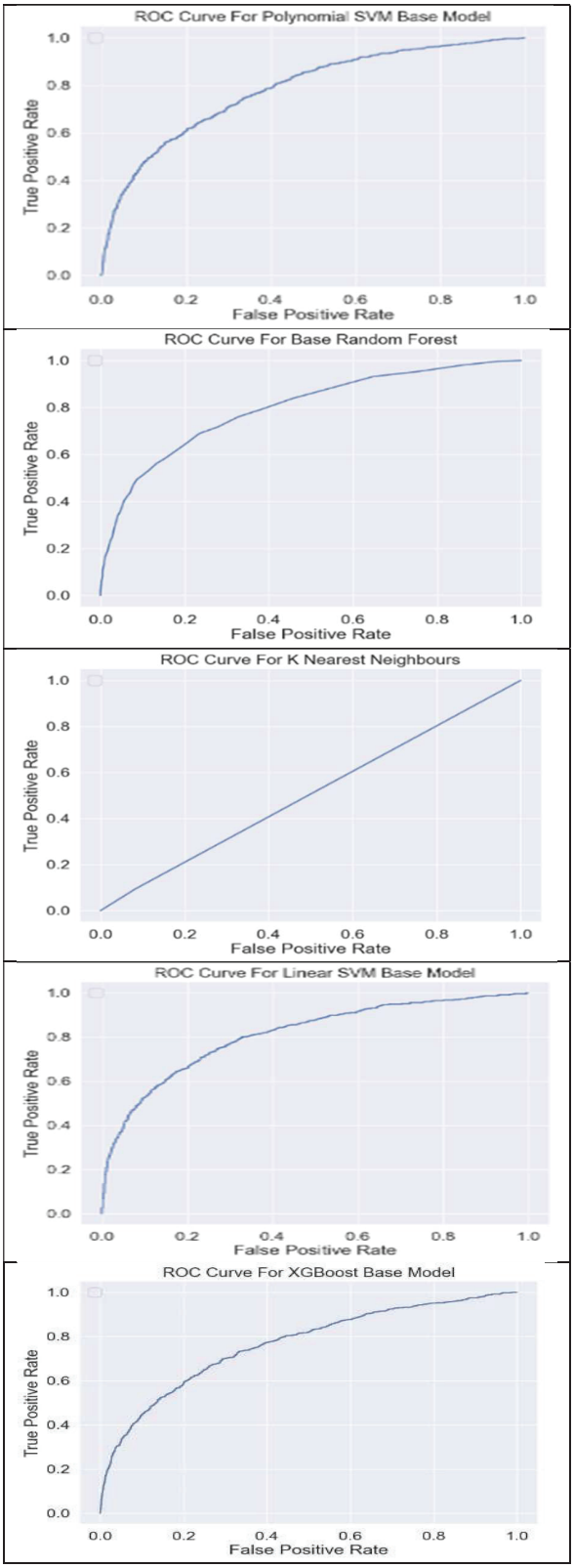

Figure 5.
Continued

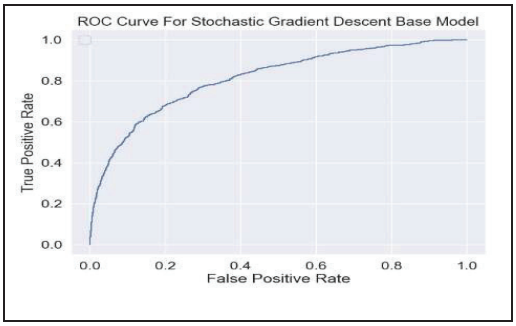

Figure 5.
ROC curves of RF, KNN, SVM, XGB, and SGD.

SVM, as shown in Table 4. We achieved the highest F1 score of recall value of 0.96 with random forest and linear SVM, as shown in Table 5. The performance of all the algorithms used can be compared using the normalized confusion matrices for binary classification and ROC curves, as shown in Figures 4 and 5, respectively.

It can be observed from Figure 4 that the base random forest correctly classified all depressed individuals as depressed (TP = 1), 96% of not depressed individuals were misclassified as depressed (FP = 0.96), and A very tiny fraction (0.12%) of truly depressed people were missed. Only 4.1% of not depressed people were correctly identified. We can understand it like this for the other models shown in Figure 4. Each point on the curve corresponds to a specific decision threshold, showing a combination of true positive rate (TPR) and false positive rate (FPR) values.

6. Conclusion

In this research work, different machine-learning algorithms were applied to the collected data. Random Forest outperformed other algorithms. We have achieved 3%, 6%, and 9% higher accuracy than [12–14]. We attained 0.99 precision using Linear SVM, which is better than [14]. Also, got a 0.96 F1 score higher than [12]. Depression continues to remain a life-degrading condition for millions. The application of machine learning can prove to be a transformative step in the healthcare industry. This work highlights the benefits of harnessing the potential of machine learning for mental health.

References

[1] V. Kaur et al., "Machine Learning for Early Detection of Child Depression: A Data-Driven Approach," 2023 2nd International Conference on Futuristic Technologies (INCOFT), Belagavi, Karnataka, India, 2023, pp. -5, doi: 10.1109/INCOFT60753.2023.10425378.

[2] M. Keerthiga, D. Abisha, P. Kalaiselvi, and S. Shenbaga Lakshmi, "Machine Learning-based Depression Prediction using Social Media Feeds," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 863–869, doi: 10.1109/ICICT57646.2023.10134427.

[3] M. H. Kabir, N. Samrat, A. Al Mahmud, R. Akter and M. Raihan, "Mental Stress Prediction from the Text of Social Media

Using Machine Learning Techniques," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1–7, doi: 10.1109/ICCCNT56998.2023.10308343.

[4] S. Nilushika Gamage and P. P. G. Dinesh Asanka, "Machine Learning Approach to Predict Mental Distress of IT Workforce in Remote Working Environments," 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 2022, pp. 211–216, doi: 10.1109/SCSE56529.2022.9905229.

[5] N. T. Singh, R. Dhiman, P. Luthra, and S. Goyal, "Predictive Analysis of Mental Stress using Machine Learning Techniques," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1269–1273, doi: 10.1109/ICCES57224.2023.10192635.

[6] A. Batabyal, V. Singh, M. K. Gourisaria and H. Das, "Sleep Stress Level Classification through Machine Learning Algorithms," 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2022, pp. 91–96, doi: 10.1109/OCIT56763.2022.00027.

[7] M. Karunakaran, J. Balusamy and K. Selvaraj, "Machine Learning Models based Mental Health Detection," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 835–842, doi: 10.1109/ICICICT54557.2022.9917622.

[8] D. Shi, X. Lu, Y. Liu, J. Yuan, T. Pan and Y. Li, "Research on Depression Recognition Using Machine Learning from Speech," 2021 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2021, pp. 52–56, doi: 10.1109/IALP54817.2021.9675271.

[9] C. A. V. Palattao, G. A. Solano, C. A. Tee and M. L. Tee, "Determining factors contributing to the psychological impact of the COVID-19 Pandemic using machine learning," 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Korea (South), 2021, pp. 219–224, doi: 10.1109/ICAIIC51459.2021.9415276.

[10] Bhakta and A. Sau "Prediction of depression among senior citizens using machine learning classifiers", International Journal of Computer Applications, vol. 144, no. 7, pp. 11–16, June 2016. DOI: 10.5120/ijca2016910429.

[11] https://wwwn.cdc.gov/nchs/nhanes/default.aspx.

[12] Chung, Jetli and Teo, Jason. (2022). Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. Applied Computational Intelligence and Soft Computing. 2022. 1–19. doi: 10.1155/2022/9970363.

[13] Abdulla, Hind, Maalouf, Maher and Jelinek, Herbert. (2023). Machine Learning for the Prediction of Depression Progression from Inflammation Markers. 2023. 1–4. doi: 10.1109/EMBC40787.2023.10340436.

[14] S. S. Malik and A. Khan, "Anxiety, Depression and Stress Prediction among College Students using Machine Learning Algorithms," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Tiruchirappalli, India, 2023, pp. 1–5, doi: 10.1109/ICEEICT56924.2023.10157693.

[15] Swati Miahra, and B. Megha Agarwal. "Diagnosis and Classification of Cancer Using Machine Learning Techniques." In 2022 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), pp. 1–5. IEEE, 2022. doi: 10.1109/SOLI57430.2022.10294965.

[16] A. Arya, R. Kumari and P. Bansal, "Predicting Depression and Mental Illness Using Machine Learning Algorithms," GV 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023, pp. 399–404, doi: 10.1109/ICCSAI59793.2023.10421262.

[17] P. Nison, P. Vuttipittayamongkol, P. Boonyapuk and K. Kemavuthanon, "A Machine Learning Approach for Depression Screening in College Students Based on Non-Clinical Information," 2023 International Conference on Cyber Management And Engineering (CyMaEn), Bangkok, Thailand, 2023, pp. 413–417, doi: 10.1109/CyMaEn57228.2023.10051001.

[18] S. Annapoorani and P. Saravanan, "From Text to Visuals: Advancements in Depression Prediction Using AI and Machine Learning Techniques," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1–6, doi: 10.1109/RMKMATE59243.2023.10369708.

[19] A. Benny, A. V. S, A. Subair, A. P. Nair and S. Thomas, "Suicidal Ideation Prediction Using Machine Learning," 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 772–776, doi: 10.1109/ICCPCT58313.2023.10245513.

[20] Swati Mishra, and Utcarsh Agarwal. "Lung cancer detection (LCD) from histopathological images using fine-tuned deep neural network." Proceedings of the International Conference on Intelligent Computing, Communication, and Information Security. Singapore: Springer Nature Singapore, 2022. doi: 10.1007/978-981-99-1373-2_19.

[21] O. S. Bankar, Y. M. Rajput, V. Kumbhar and T. P. Singh, "Machine Learning Applications in Depression Research: A Comprehensive Review and Analysis," 2023 International Conference on Integration of Computational Intelligent System (ICICIS), Pune, India, 2023, pp. 1–6, doi: 10.1109/ICICIS56802.2023.10430263.

[22] A. M. Chekroud, R. J. Zotti, Z. Shehzad et al., "Cross-trial prediction of treatment outcome in depression: a machine learning approach," Ae Lancet Psychiatry, vol. 3, no. 3, pp. 243–250, 2016.

[23] S. Rudenstine, K. McNeal, T. Schulder, C. K. Ettman, M. Hernandez, K. Gvozdieva, et al., "Depression and anxiety during the covid- 19 pandemic in an urban low-income public university sample", *Journal of Traumatic Stress*, vol. 34, no. 1, pp. 12–22, 2021.

[24] Long Xu, Xin Shu, and Jian Shu, "Research on Depression Tendency Detection Based on Image and Text Fusion", *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD).*, 2022.

[25] Md. Mehedi Hassan, Md. Asif Rakib Khan, Khan Kamrul Islam, Md. Mahedi Hassan and M M Fazle Rabbi, "Depression Detection System with Statistical Analysis and Data Mining Approaches", *2021 International Conference on Science & Contemporary Technologies (ICSCT).*, 2021.

[26] K. A. G. a. N. Palanichamy, "Depression Detection Using Machine Learning Techniques on Twitter Data", *International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 960–966, 2021.

[27] M. R. A. R. J. M. A. R. A. S. P. M. U. Amna Amanat, "Deep Learning for Depression Detection from Textual Data", *electronics*, no. February 7, pp. 1–13, 2022.

[28] B. S. I. J. E. J. A. N. J. A. P. Zannatun Nayem Vasha, "Depression detection in social media comments data using machine learning algorithms", *Bulletin of Electrical Engineering and Informatics*, no. August 6, pp. 987–995, 2022.

[29] "Stress detection using natural language processing and machine learning over social interactions", *Joourof Big Data*, pp. 1–24, 2022.

[30] "Early Depression Detection from Social Network Using Deep Learning Techniques", *vol. IEEE Region 10 Symposium (TENSYMP)*, no. June 7, pp. 823–826, 2022.

[31] S. Mishra and M. Agarwal, "Skin Cancer Classifier: Performance Enhancement Using Deep Learning Models," 2025 10th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2025, pp. 721–725, doi: 10.1109/ICSC64553.2025.10969043.

[32] S. Mishra and D. Srivastava, "Employing Machine Learning Techniques for Depression Prediction," 2024 3rd International Conference for Advancement in Technology (ICONAT), GOA, India, 2024, pp. 1–4, doi: 10.1109/ICONAT61936.2024.10775113.

Biographies

Swati Mishra received her B. Tech degree in Electronics and Instrumentation Engineering from Uttar Pradesh Technical University, Lucknow, in 2005, and M. Tech. in Control and Instrumentation from Rajasthan Technical University, Kota, in 2014. Currently pursuing a PhD. from Jaypee Institute of Information and Technology, Noida. Held and similar information for other professional societies. She has been working as an Assistant Professor at JSS Academy of Technical Education and Technology, Noida, since 2008. Her research interests include Biomedical Image Processing, Artificial Intelligence, and Computer Vision. She has published papers in International Conferences and book chapters.

Divyanshi Srivastava, a final-year undergraduate student pursuing a Bachelor of Technology in Electrical and Electronics Engineering from JSS Academy of Technical Education, Noida. She is a passionate and motivated technologist, particularly in the fields of Computer Vision, Artificial Intelligence, and Image Processing.