Dean&Francis
ISSN 2959-6157

Prediction of Depression and Anxiety based on Machine Learning

Yuqiu Tian[1, *]

[1]School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China

*Corresponding author: tyuqiu00@gmail.com

Abstract:

Depression and anxiety are prevalent and often comorbid mental health disorders, necessitating accurate prediction and timely intervention to mitigate their impact. This study focuses on leveraging advanced machine learning techniques to predict the levels of depression and anxiety, while also exploring the intricate correlation between these two conditions. Using a dataset comprising depression and anxiety scores, several Machine Learning models, including the SVM, KNN, XGBoost and CatBoost were employed to develop predictive models. The experimental results revealed that CatBoost outperformed other models, achieving an accuracy of 98.9%, followed closely by SVM. Additionally, Pearson and Spearman correlation analyses demonstrated a strong relationship between depression and anxiety scores, with coefficients of 0.6261 and 0.6208, respectively. These findings underscore machine learning's capacity to forecast mental health issues and highlight the significant correlation between depression and anxiety, providing a robust basis for developing more effective and timely early intervention strategies to improve mental health outcomes.

Keywords: Depression; anxiety; machine learning; correlation.

1. Introduction

There has been significant advancement in the use of machine learning in mental health, particularly in predicting depression and anxiety. Machine learning for early diagnosis and intervention has become a vital field of research as these mental health disorders spread throughout the world. This chapter highlights the recent advancements in this field, examining various methods used for predicting depression and anxiety, and discussing their benefits, challenges, and future research directions.

Shatte et al. conducted an extensive literature review, demonstrating the considerable effectiveness of machine learning for psychological well-being [1]. However, while machine learning technologies have shown success in disease detection, their application in other areas warrants further investigation. Building on this foundation, Ahmed et al. developed a model based on standard psychological assessments and machine learning algorithms specifically aimed at predicting varying severity levels of depression and anxiety [2]. Their study evaluated five AI al-

ISSN 2959-6157

gorithms and the Convolutional Neural Network (CNN) demonstrated exceptional accuracy, achieving 96.8% and 96% in predicting depression and anxiety, respectively, among women in Bangladesh. Following this, Nemesure et al. concentrated on using Electronic Health Records (EHRs) to predict Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD) [3]. They designed a novel machine learning pipeline that integrated multiple algorithms, including deep learning, to analyze data from 4,184 university students. The prediction AUCs for GAD and MDD were 0.73 and 0.67, respectively, indicating moderate predictive performance.

Relatedly, Aggarwal and Goyal expanded the scope of machine learning applications, especially in anticipating the COVID-19 pandemic-related anxiety and depression experienced by online gamers [4]. Their study combined gaming behavior data with self-esteem measurements and applied four machine learning algorithms to data from over 3,500 internet gamers. The Decision Tree classifier achieved the highest accuracy (100%) in predicting GAD and 84.71% accuracy in evaluating the Satisfaction with Life Scale (SWL). The same year, Kilaskar et al. further enriched the discussion by evaluating the accuracy of different algorithms in predicting depression based on user social behavior and data, emphasizing the significance of unconventional data sources [5]. Chung and Teo conducted a comprehensive review that explored the various applications of machine learning techniques within the field of mental health [6]. They investigated the efficacy of machine learning in treating mental health conditions such bipolar disorder, schizophrenia, and depression by examining thirty pertinent research. This aligned with Kilaskar's findings, which underscored the importance of feature selection and data processing.

Nickson et al. reviewed studies from 2011 onward on using EHRs and machine learning technology to predict depression [7]. Despite difficulties with standardization and generalization, they found that integrating machine learning with EHRs had a great deal of potential for predicting depression based on their analysis of nineteen studies. In order to increase clinical application, they recommended that future research concentrate more on enhancing the generalizability and interpretability of models. Tian et al. investigated how social support and resilience affected depression and anxiety during the epidemic [8]. Analyzing data from 29,841 participants, Tian and his colleagues used an XGBoost model to identify key variables affecting mental health and classified participants into healthy control, depression, anxiety, and comorbid groups. Jha et al. employed the DASS42 questionnaire and the WESAD dataset to predict depression and anxiety using four categories of machine learning algorithms [9]. Jha et al. indicated that combining questionnaire data with wearable device readings allowed for more accurate detection of depression and anxiety. This outcome resonated with Tian et al.'s findings using the XGBoost model, further validating machine learning's potential in mental health prediction. Tasnim et al. introduced a CNN and VGG-19 features-based model for predicting the severity of depression, anxiety, and stress through speech features [10]. Using a longitudinal dataset, they analyzed acoustic features in speech samples and found that their model accurately predicted the severity of mental health issues. The study highlighted the advantages of using speech features for data privacy and language independence.

In conclusion, a systematic analysis of recent literature indicates that machine learning technology holds great promise in predicting depression and anxiety disorders. Various research methods and datasets offer diverse solutions, though challenges remain in data standardization, model interpretability, and generalizability. The research aims to improve mental health prediction accuracy and dependability by utilizing these computational tools, which will ultimately lead to more successful early detection and intervention measures, as well as to investigate the relationship between these two conditions for the purpose of facilitating early intervention in depression.

2. Methods

2.1 Data Source

The Depression Anxiety Stress Scales (DASS) online version was used to collect data for this study between 2017 and 2019. The data was sourced with Kaggle. The Anxiety scale gauges situational anxiety and autonomic arousal, whereas the Depression scale evaluates dysphoria, hopelessness, and associated symptoms. On 4-point rating scales, participants rate their experiences from the previous week; scores are added together for each dimension. The DASS scales are useful for assessing present circumstances and treatment-related changes since they have a high level of internal consistency and can distinguish across states.

2.2 Feature Extraction and Filtration

In order to assure a more characteristic model, this paper mapped the emotional qualities with various personalities after filtering the acquired data to only extract the aspects that are necessary to determine an individual's emotional status (Table 1). The feature "major" was removed due to being a text feature with over 50% missing values and low correlation with the target variable.

The information has been cleansed and prepared in advance of machine learning model's further training and testing. The interquartile range was initially determined in the study, after which the data was normalized, and any outliers were cleaned up.

Table 1. Descriptive Summary of Dataset Features

Feature Name	Type	Range
Q1A-Q42A	Category	1 = Didn't really apply to me, 2 = Pertained to me in part, or occasionally, 3 = Applied to me a good portion of the time, or to a significant extent, 4 = Very often, or most of the time, applied to me
TIPI1-TIPI10	Category	1 = A sharp disagreement, 2 = A moderate disagreement, 3 = A slight disagreement, 4 = Not in agreement or disagreement, 5 = Slightly agree, 6 = Moderate agreement, 7 = Strong agreement
Education	Category	1 = Not much higher education, 2 = High school, 3 = University degree, 4 = Graduate degree
Gender	Category	1=Male, 2=Female, 3=Other
Age	Integer	13-80
Religion	Category	1= Christian (Protestant), 2 = Christian (Other), 3 = Hindu, 4 = Sikh, 5 = Agnostic, 6 = Atheist, 7 = Jewish, 8 = Buddhist, 9 = Christian (Catholic), 10 = Christian (Mormon), 11= Muslim, 12 = Other
Orientation	Category	1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other
Race	Category	10 = Arab, 20 = Asian, 30 = Native American, 40 = White, 50 = Black, 60 = Indigenous Australian, 70 = Other
Married	Category	1=Never married, 2=Currently married, 3=Previously married
Familiysize	Integer	0-19, 21,23,24,26
DS	Integer	0-42
AS	Integer	0-42
DL	Category	1 = Normal, 2 = Mild, 3 = Moderate, 4 = Severe, 5 = Extremely Severe
AL	Category	1 = Normal, 2 = Mild, 3 = Moderate, 4 = Severe, 5 = Extremely Severe

2.3 Model Introduction

2.3.1 Support vector machine

Support Vector Machine (SVM), one of the most widely utilized supervised learning models for classification tasks, operates by identifying the optimal hyperplane within a feature space. By using kernel functions to translate inputs into higher-dimensional spaces where they become linearly separable, SVM is able to handle non-linear data. It is robust for high-dimensional data and is effective in applications like image and text classification.

2.3.2 K-nearest neighbors (KNN)

This approach relies on proximity to existing data points to make accurate predictions. The distance metric, commonly Euclidean, determines neighbor proximity. KNN is intuitive, requires no training phase, and performs well with small datasets but may struggle with high-dimensional or large-scale data due to increased computational complexity.

2.3.3 XGBoost

A group of decision trees is constructed with the scalable, high-performance gradient boosting framework XGBoost in order to maximize prediction accuracy. In order to increase efficiency and speed, it makes use of sophisticated strategies like tree pruning, parallel processing, and handling missing data. A regularized objective function is also used to prevent overfitting. XGBoost is particularly effective for structured data and has become a popular choice in machine learning competitions due to its robust performance and flexibility in both classification and regression tasks.

2.3.4 CatBoost

CatBoost is a gradient boosting algorithm designed for high-performance processing of categorical data. It employs ordered boosting and efficient encoding techniques to manage categorical features without extensive pre-pro-

ISSN 2959-6157

cessing. CatBoost mitigates overfitting through unique regularization strategies and manages missing data and outliers effectively. It is highly efficient, supports both classification and regression tasks, and is particularly suited for applications involving large-scale, structured datasets.

3. Results and Discussion

By analyzing their predictive capabilities, the research seeks to determine how effectively these algorithms can contribute to the early identification of individuals at risk, ultimately aiding in more proactive mental health management strategies. The Ten Item Personality Inventory, the DASS-42 questionnaire answers, and other participant demographic data are combined into one dataset. The following four machine learning models were evaluated and compared for their predictive power. The two kinds of correlation coefficient were also computed to examine the relationship between depression and anxiety.

3.1 Data Visualization

To provide a more intuitive visualization and comparison of the number of individuals across different depression levels by age, education, gender, orientation, religion, race, marital status, and family size, this study utilizes grouped bar charts for data visualization. The following figure illustrates that, regardless of the level of depression, especially in cases of extremely severe depression, most participants are aged between 13 and 30. This suggests that depressive tendencies are more pronounced among adolescents and young adults (Fig. 1).

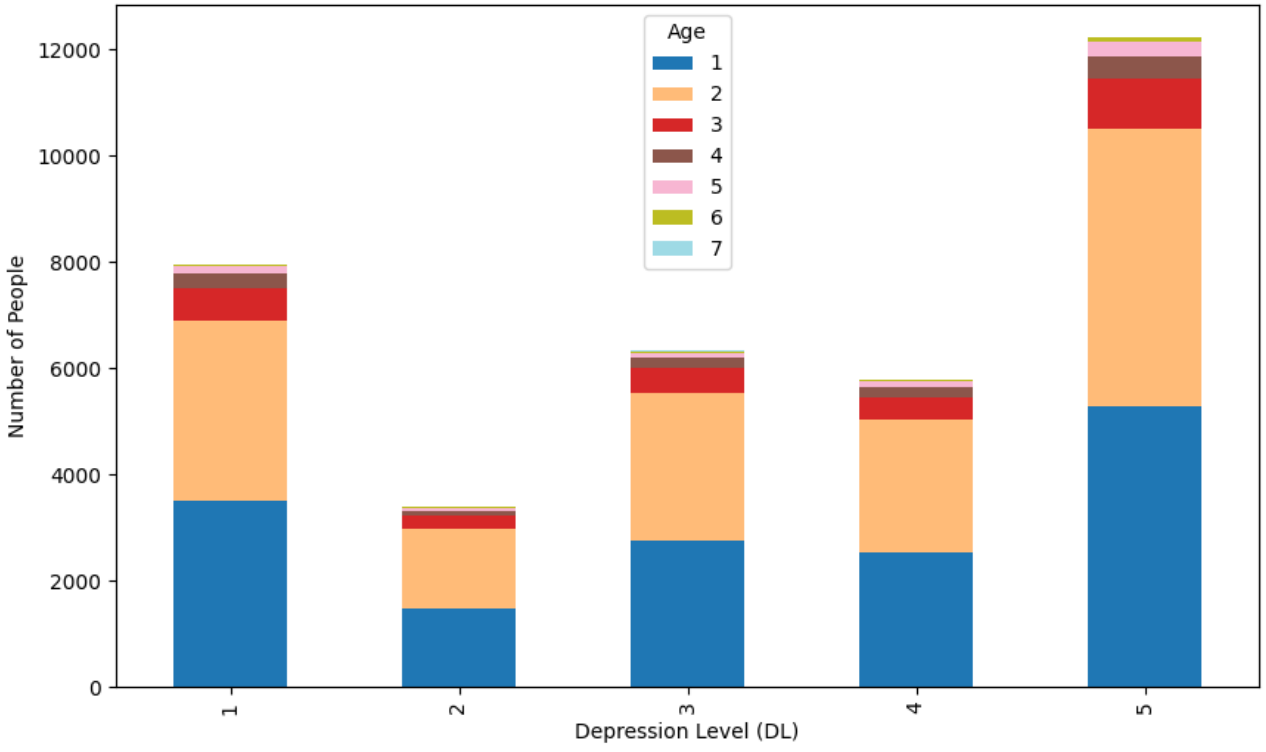

Fig. 1 Depression Level Distribution across Different Age

Additionally, this study reveals that participants with severe and extremely severe depression tend to have an education level concentrated at the high school and university levels, which corresponds to the observed age distribution trend (Fig. 2).

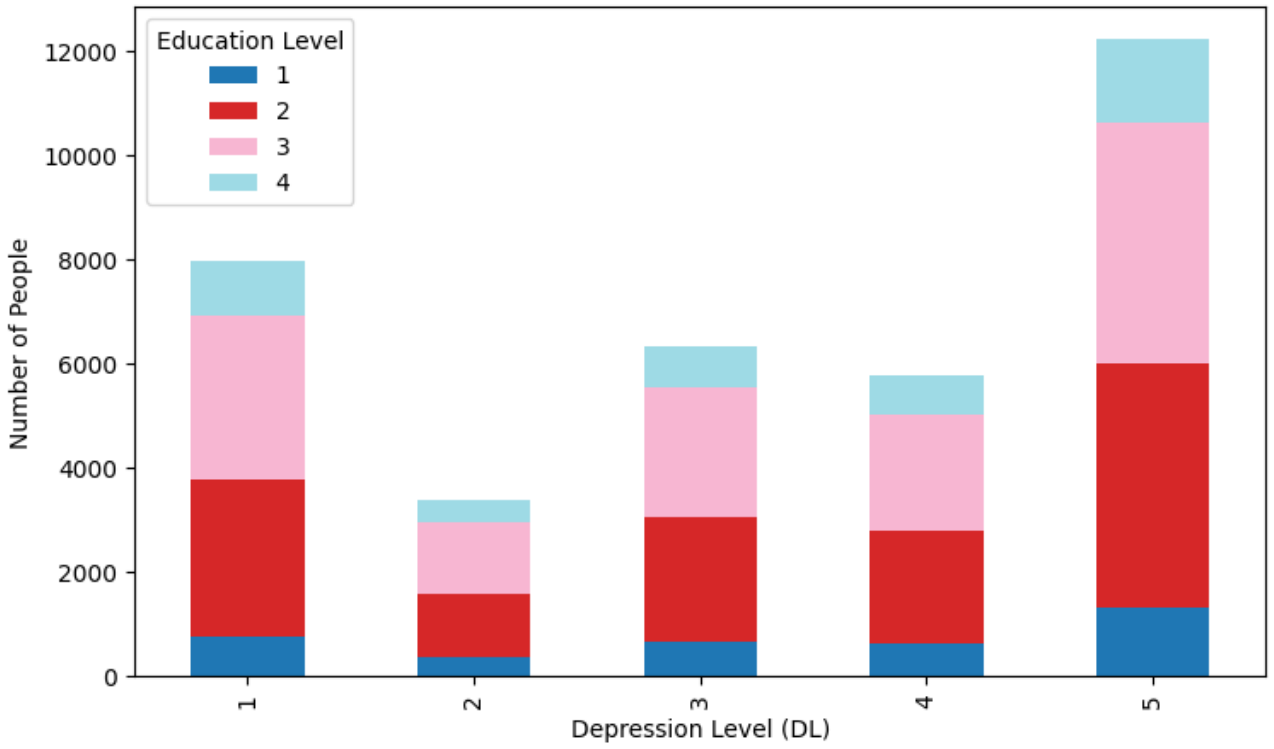

Fig. 2 Depression Level Distribution across Education Level

Similarly, the majority of participants with depressive tendencies or severe depression are female, indicating that women are more susceptible to depression compared to men and individuals of other genders (Fig. 3).

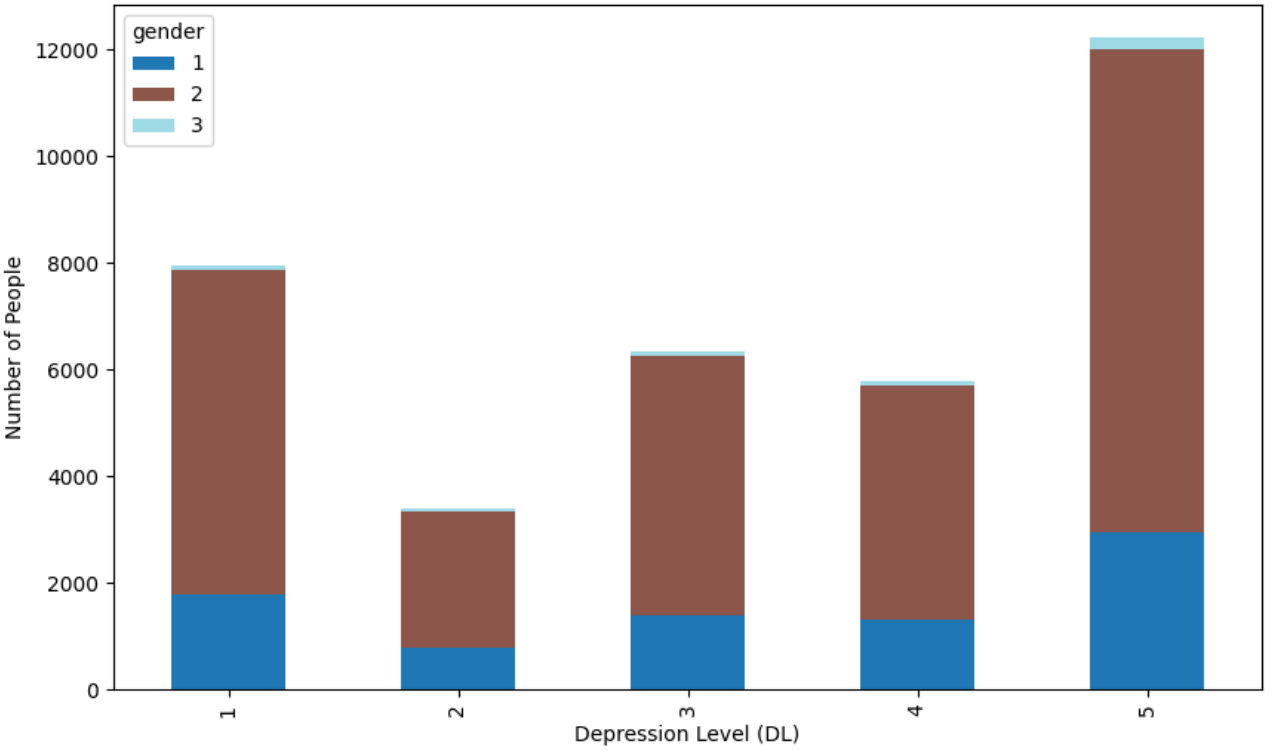

Fig. 3 Depression Level Distribution across Different Gender

The study also shows that sexual orientation has a significant relationship with the level of depression. According

ISSN 2959-6157

to the distribution of participants by orientation, aside from heterosexual individuals, who constitute the majority, bisexual and other sexual minority groups are more prone to experiencing severe depressive tendencies (Fig. 4).

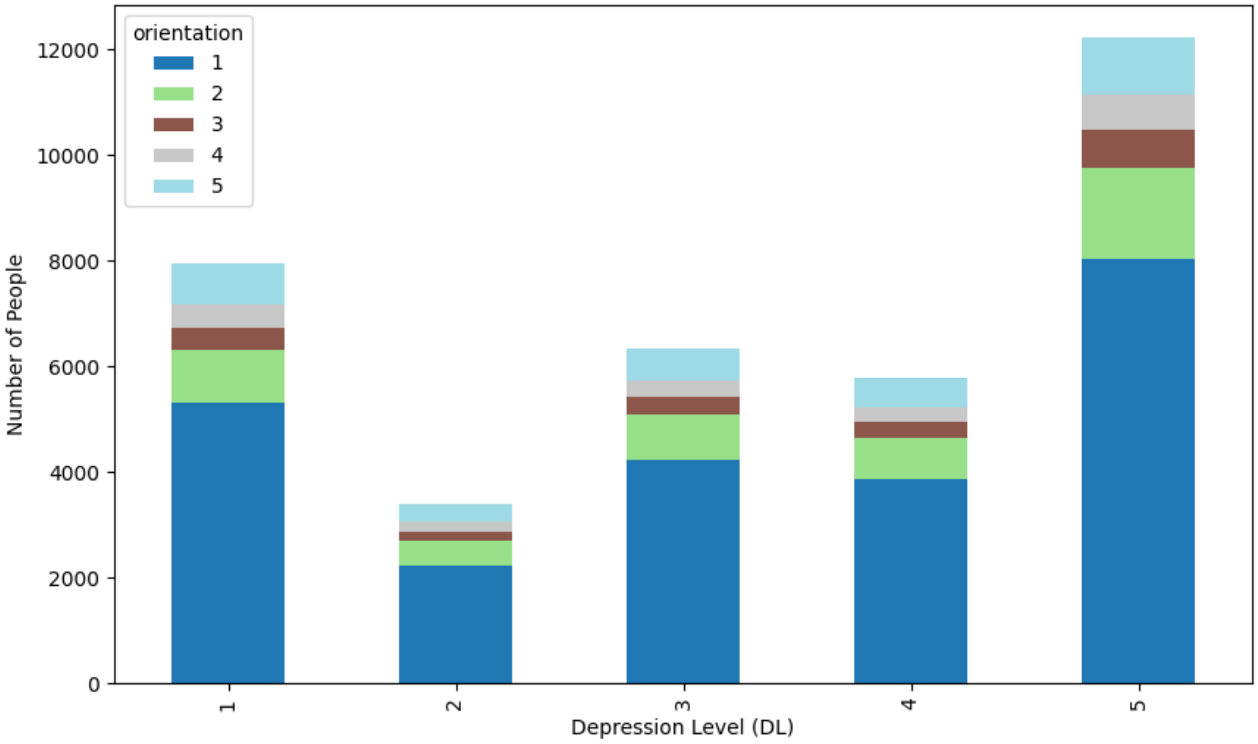

Fig. 4 Depression Level Distribution across Different Orientation

3.2 Prediction Results

Based on the given dataset, four machine learning models were used in this work to predict anxiety and depression. A performance evaluation of each model's classification was conducted, concentrating on important metric. Below is a detailed discussion of the analyses' findings, outlining each model's advantages and disadvantages as well as how well it predicts the desired results overall.

3.2.1 SVM model results

The overall performance of the SVM model is exemplary, with an accuracy of 0.9785, indicating a highly precise classification across all categories. Additionally, the weighted averages for these metrics, 0.9784, 0.9785, and 0.9785, further underscore the model's robustness, accounting for the varying sample sizes within each category (Table 2). These results confirm the model's strong predictive capabilities and its suitability for accurate and efficient classification across a diverse dataset.

Table 2. Support Vector Machine Performance Parameters

	Precision	Recall	F1-score	Support
1	0.9875	0.9856	0.9865	1599
2	0.9495	0.9430	0.9462	737
3	0.9668	0.9707	0.9688	1230
4	0.9672	0.9663	0.9667	1128
5	0.9923	0.9939	0.9931	2457
Accuracy			0.9785	7151
Macro avg	0.9726	0.9719	0.9723	7151
Weighted avg	0.9784	0.9785	0.9785	7151

3.2.2 KNN results

The KNN model exhibits a considerably lower overall performance compared to the SVM, with an accuracy of 0.7591, indicating less precise classification across all categories (Table 3).

Table 3. K-Nearest Neighbors Performance Parameters

	Precision	Recall	F1-score	Support
1	0.8210	0.9325	0.8732	1599
2	0.4713	0.3338	0.3908	737
3	0.5839	0.7187	0.6443	1230
4	0.6398	0.5133	0.5696	1128
5	0.9307	0.9068	0.9186	2457
Accuracy			0.7591	7151
Macro avg	0.6893	0.6810	0.6793	7151
Weighted avg	0.7533	0.7591	0.7518	7151

3.2.3 XGBoost results

The XGBoost model demonstrates strong overall performance, with an accuracy of 0.9646, indicating highly precise classification across all categories (Table 4). Compared to other models like KNN, XGBoost offers superior predictive capabilities, making it a reliable choice for tasks requiring high classification accuracy in diverse datasets.

Table 4. XGBoost Performance Parameters

	Precision	Recall	F1-score	Support
1	0.9766	0.9906	0.9835	1599
2	0.9468	0.8697	0.9066	737
3	0.9207	0.9537	0.9369	1230
4	0.9476	0.9450	0.9463	1128
5	0.9923	0.9906	0.9914	2457
Accuracy			0.9646	7151
Macro avg	0.9568	0.9499	0.9530	7151
Weighted avg	0.9647	0.9646	0.9644	7151

3.2.4 CatBoost results

The CatBoost model demonstrates outstanding performance, achieving an accuracy of 0.9890, which indicates exceptionally precise classification across all categories (Table 5). Compared to other models such as KNN and even XGBoost, CatBoost offers superior predictive capabilities, making it an excellent choice for tasks that require high classification accuracy and reliability in diverse datasets.

Table 5. CatBoost Performance Parameters

	Precision	Recall	F1-score	Support
1	0.9925	0.9987	0.9956	1599
2	0.9902	0.9620	0.9759	737
3	0.9760	0.9902	0.9831	1230
4	0.9813	0.9796	0.9805	1128
5	0.9963	0.9943	0.9953	2457
Accuracy			0.9890	7151

ISSN 2959-6157

Macro avg	0.9873	0.9850	0.9861	7151
Weighted avg	0.9890	0.9890	0.9889	7151

3.2.5 Comparison results

CatBoost exhibited the highest predictive performance for depression prediction, achieving an accuracy of 98.9%. This superior performance can be attributed to CatBoost's advanced gradient boosting algorithms and its ability to effectively handle categorical features and imbalances in the dataset. In contrast, the SVM model demonstrated impressive performance as well but fell slightly short of CatBoost. The KNN model, while achieving a respectable accuracy of 75.9%, performed the least effectively among the models. This lower performance is likely due to KNN's reliance on distance metrics, which can be less effective in high-dimensional or imbalanced datasets. The experimental results are summarized in the accompanying bar chart (Fig. 5).

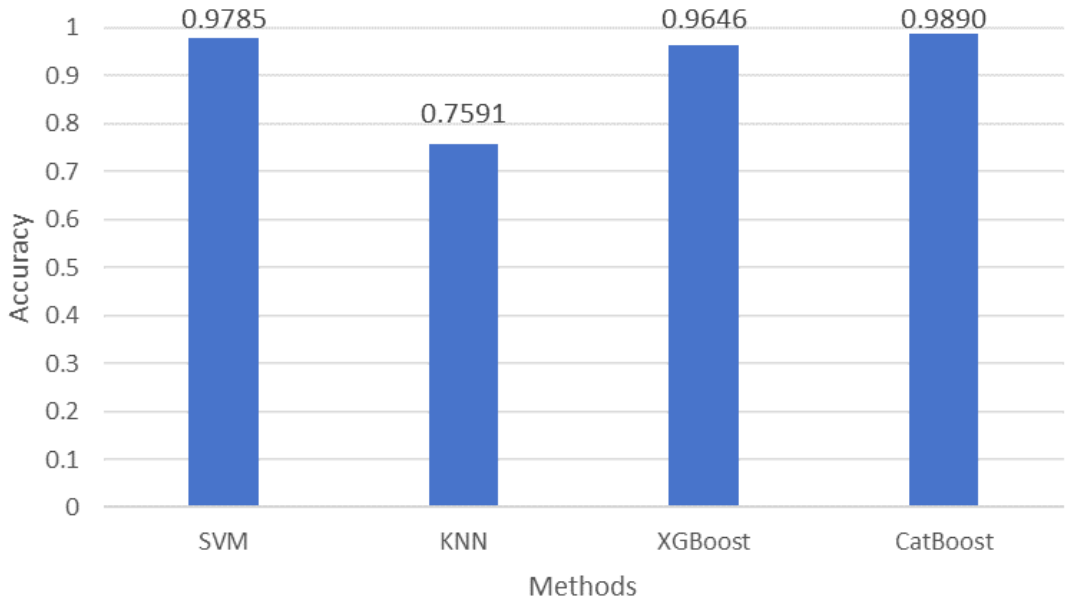

Fig. 5 Prediction Accuracy of Four Machine Learning Models

3.3 Correlation between Depression and Anxiety

The Pearson Correlation Coefficient (Table 6) and Spearman's Rank Correlation Coefficient (Table 7) were used to evaluate the link between the depression level (DL) and anxiety level (AL). The analysis yielded coefficients of 0.6261 and 0.6208, respectively, indicating a strong correlation between the two variables.

Table 6. Pearson Correlation Coefficient

	DL
AL	0.6261**

Table 7. Spearman's Rank Correlation Coefficient

	DL
AL	0.6208**

Practically speaking, this high link suggests that anxiety and depression tend to rise in tandem with rising levels of each other. This finding highlights the interconnected nature of depression and anxiety, supporting the need for

integrated approaches in research and treatment for these mental health conditions.

4. Conclusion

Through extensive experimentation and in-depth analysis, CatBoost was identified as the most efficient and high-performing model, demonstrating remarkable accuracy. It achieved a notable 98.9% accuracy on the dataset, surpassing other models in its ability to accurately predict outcomes. This underscores CatBoost's potential as a powerful tool for predictive modeling. Support Vector Machine (SVM) also performed well, closely following CatBoost, while K-Nearest Neighbors (KNN) exhibited the least ideal performance among the tested models but still maintained a respectable accuracy of 75.9%. Furthermore, using the Pearson Correlation Coefficient and Spearman's Rank Correlation Coefficient, the association between depression levels (DL) and anxiety levels (AL) was carefully assessed. The resulting coefficients indicate a strong correlation between these two psychological metrics, which underscores the interrelated nature of depression and anxiety.

The findings of this study highlight the significant correlation between anxiety and depression, emphasizing the close relationship between these two mental health conditions. Additionally, the results demonstrate the encouraging possibilities of models for machine learning as effective tools for predicting and understanding this interconnection, paving the way for improved diagnostic and intervention strategies, in particular CatBoost, to predict mental health issues with accuracy. These discoveries may help design more focused and successful early intervention programs for anxiety and depression. However, future work should address the limitations of the current models and explore additional variables that might further enhance predictive accuracy and clinical applicability.

References

[1] Shatte A B R, Hutchinson D M, Teague S J. Machine learning in mental health: a scoping review of methods and applications. Psychological medicine, 2019, 49(9): 1426-1448.

[2] Ullas T R, Begom M, Ahmed A, Sultana R. A Machine Learning Approach to detect Depression and Anxiety using Supervised Learning. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020: 1-6.

[3] Nemesure M D, Heinz M V, Huang R, Jacobson N C. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Scientific reports, 2021, 11(1): 1980.

[4] Aggarwal R, Goyal A. Anxiety and Depression Detection using Machine Learning. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 2022: 141-149.

[5] Kilaskar M, Saindane N, Ansari N, et al. Machine Learning Algorithms for Analysis and Prediction of Depression. SN COMPUT. SCI, 2022, 3: 103.

[6] Chung Jetli, Teo Jason. Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. Applied Computational Intelligence and Soft Computing, 2022: 1-19.

[7] Nickson D, Meyer C, Walasek L, et al. Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. BMC Med Inform Decis Mak, 2023, 23: 271.

[8] Tian Z, Qu W, Zhao Y, Zhu X, Wang Z, Tan Y, Jiang R, Tan S. Predicting depression and anxiety of Chinese population during COVID-19 in psychological evaluation data by XGBoost. Journal of affective disorders, 2023, 323: 417-425.

[9] Jha A, Abirami M S, Kumar V. Predictive Model for Depression and Anxiety Using Machine Learning Algorithms. Deep Sciences for Computing and Communications. IconDeepCom 2022. Communications in Computer and Information Science, 2023, 1719.

[10] Tasnim M, Diaz Ramos R E, Stroulia E, Trejo L A. A Machine-Learning Model for Detecting Depression, Anxiety, and Stress from Speech. ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024: 7085-7089.