International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 1
ISSN: 2088-5334

Cluster Ensemble Method and Convolution Neural Network Model for Predicting Mental Illness

Ananthapadmanabha M V [a], Dhanesh Kumar A C [a], Sabariraju S [a], Eswar M [a], Senthilkumar Mathi [a,*]

[a] *Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, Amrita Vishwa Vidyapeetham, India*
*Corresponding author: *m_senthil@cb.amrita.edu*

Abstract **—One in every four individuals has a diagnosable mental disease in a year. Around 20% of children and adolescents have a mental health condition and often ignore it. It is found that 93% of youth use social media to communicate and engage, as it reflects their emotions, moods, and thoughts. As a result, machine learning algorithms may anticipate people's moods and emotions based on their postings and comments. On the other hand, psychometric tests use questions to determine how individuals think, feel, behave, and react. It is necessary to investigate a hybrid approach for identifying people's mental illness by combining social media inputs and psychometric tests, especially during a pandemic. The hybrid approach can combine the results from both the models to reflect on the user's digital & non-digital reactions to certain sensitive situations to determine their mental state. Hence, the present paper aims to develop a web framework that can forecast the emergence of mental illness in the future based on data from social media comments and real-time data from psychometric tests using machine learning algorithms. The proposed work includes the cluster ensemble method for social media posts and a convolution neural network model for psychometric tests. This model predicts mental illness with an accuracy of 87.05 percent. The individual can use this result to take the required precautions by visiting a psychologist.**

Keywords— **Machine learning; classification; deep learning; feature extraction; neural network.**

Manuscript received 26 Mar. 2022; revised 20 Jul. 2022; accepted 19 Sep. 2022. Date of publication 28 Feb. 2023.

I. INTRODUCTION

A prevalent issue in the medical field is medical health problems. However, identifying mental health problems is time-consuming, costly, and often delayed. Early detection of mental health problems is essential to understand mental health conditions better and providing better patient care [1]. Artificial intelligence and machine learning technologies are being used to increase our knowledge of mental health disorders and to aid mental health clinicians in making better clinical decisions, thanks to the increasing availability of data regarding an individual's mental health status. Machine learning and deep learning, which transform input through layers of nonlinear computational processing units, give a new paradigm for successfully gaining information from complicated data as one of the most recent advancements in artificial intelligence [2].

Deep learning algorithms have improved performance in various data-rich applications, including healthcare. Mental illness is a sickness that affects a person's thinking, emotions, or behavior (or all three) and has been linked to physical health problems. It influences a person's relaxed state of mind. Social media has started to creep most of our time [3]. Hence, this paper aims to develop a framework to predict mental illness using hybrid social media posts and psychometric tests.

II. MATERIALS AND METHOD

A. Related Works

Thorstad and Wolff [1] focus on whether people's everyday language contains sufficient signals to predict the future occurrence of mental illness. Data samples were collected from clinical subreddits (directly linked with mental illness) and non-clinical subreddits (not directly connected with mental illness). They were used to predict mental illness. Subreddits related to non-clinical subreddits were used to predict future posting to clinical subreddits. Models trained on clinical subreddits learned to focus on words indicating disorder-specific symptoms. It is also trained to anticipate future mental illness and learned to focus on words meaning life stress, suggesting that features predictive of mental illness may change over time.

Costello et al. [2] mainly focus on accounts followed by Twitter users. It examines how psychological traits relevant to a person's mental health and well-being can be inferred from their Twitter friends. This model is apt for users trying to hide their illnesses by manipulating their tweets. Though the psychological meaning of Twitter friends is perhaps less immediately apparent than the psychological sense of tweets, sometimes accounts users follow unknowingly disclose sensitive information about their psychological traits. Many people are living in a mobility environment [7]. Yan et al. [3] aim to build a model using a convolution neural network (CNN) for the automatic detection and classification of personal mental illness at the workplace using the emotional labor and mental health questionnaire to help the staff assess their mental health on their own [4], [5]. The classification is achieved using Rasch analysis, and the work classifies less confident responses as outliers [6].

CNN is used for building predictive modeling [8], [9], [10]. The study also explores the potential benefits that improve effective classifications of true or false mental illness. Since many features can be extracted from speech, effective data selection is critical for success, given the limited training data available. Demiroglu et al. [11] focus on two novel contributions that exploit multiple language databases for acoustic feature selection. The multi-lingual method effectively selected better features than the baseline algorithms, significantly improving the depression assessment accuracy. The second contribution focuses on extracting text-based features for depression assessment and using a novel algorithm to fuse the text and speech-based classifiers, which further boosted the performance [12].

Silveira et al. [13] analyzed how discussions in Reddit communities related to mental disorders can help improve the health conditions of their users. The emotional tone of users' relationships between user interactions and state changes is found. First, the author's negative posts indicate that their emotional state can improve due to social support. Second, models are built based on state-of-the-art text embedding techniques and recurrent neural networks to predict shifts in emotional tone. Accuracy of predicting the users' reactions to the interactions experienced on these platforms is more viable, considering the examples that illustrate that the models capture the effects of comments on the emotional tone.

Nasir et al. [14] proposed a methodology for identifying mental illness via their communication on social media networks. Social media is an excellent source of communication and interaction among people. Hence, this approach is used to build a classification model. Cosic et al. [15] analyzed how the coronavirus disease 2019 (COVID-19) pandemic and its immediate results severely threaten mental health. Health care workers (HCW) may develop elevated anxiety, depression, post-traumatic stress disorder, or suicidal behaviours. Therefore, this article addresses the prevention of HCW mental health disorders by early predicting mental health disorders in health care workers.

The hybrid approach of adapting social media comments and psychometric tests can improve the integrity of the data used for training the model. Social network data and data collected from questionnaires contribute to the prediction of mental health, and an underexploited link exists between them. The hybrid approach can improve the strength of the data, which is very much needed for any machine learning algorithm in mental health diagnostics. Feature extraction from questionnaire data could boost the efficiency of training classification models for social network data [16].

B. Proposed Method - Dataset Preparation and Pre-processing

Twint package is used to scrape data from Twitter with keywords like depression, anxiety, and bipolar, which point to severe mental illness [17]. The data obtained finally is unlabeled keyword-based web-scraped data. Data is preprocessed by first cleaning the data by removing stop words like is, a, and the from the dataset, and then the data is made free of punctuations, username, and HTML tags. N-grams are used to group words based on the factor and bag-of-words to describe the occurrence of words within a document. As emojis also play a vital role in deciding the sentiment of a tweet, the emojis are converted to text. Their emotion is analyzed by using the text2emotion package. Using lemmatization and stemming techniques, the words with different tenses are normalized to their root words.

Feature extraction uses term frequency and inverse term frequency to extract the crucial words from tweets, contributing more to the sentiment. Then cosine similarity is used to remove the tweets similar to other tweets. Word2Vec is then used to convert the words into vectors, and the most frequently occurring words are found from it. Word2Vec converts the words into high-dimension, so to reduce it to two-dimension, the t-distributed stochastic neighbor embedding technique is used so that when training the model, the complexity of the input reduces. The preprocessing of psychometric data starts with removing missing values from the data and then checking for outliers. The factor analysis is performed to reduce the dimension of the dataset. Perform Rasch analysis to validate the data in the responses non-linearly.

Jamovi software is used to implement the exploratory factor analysis. Kaiser-Mayer-Olkin (KMO) and Bartlett's test is performed to test the adequacy of variables. KMO finds the variance proportion of the variables must be more than 0.5. Bartlett's test to assess the correlation in the variables must be less than 0.05. Then, the commonality is measured, and the verified generated factor matrix with factor loadings greater than 0.5 and eigenvalue more significant than one.

The data samples are taken for each analysis, and several checks are used as in [6] to validate the dataset. Reliability analysis for persons and items, usually the coefficient is greater than 0.8. Fit statistics for persons and items must be between (0.8 to 1.2). Unidimentionality and local independence must be less than 0.3. It is necessary to do preprocessing to make the data fall into these criteria.

C. Social Media Comments Model

Unsupervised techniques such as clustering can be adapted for training the model since the data is unlabeled. The single clustered approach can make biased predictions; hence, the cluster ensemble technique is adapted and investigated in [18] and [19]. The cluster similarity technique combines various clustered output data into one cluster. Using ordering points to identify the clustering structure (OPTICS) algorithm and silhouette score parameters for clustering techniques [20].

Then, the density-based clustering is performed like density-based spatial clustering of applications with noise (DBSCAN) and distance-based clustering like K-Means. They can be combined to get high accuracy clustered data using cluster ensembles. Calinski-Harabasz index is used to find if a cluster is dense and well separated. Dunn index is used to see if a random cluster structure is present because of noisy data in the dataset.

D. Psychometric test model

A multi-class CNN classifier is the technique used for classification. The dataset is divided into five main classes. Initially, the train test with the split ratio of (70:30) is used. The layers used are input, convolution, pooling, and output. The filters are used in the convolution layer, and the feature map is generated for each filter used to create the pool layers. Then, the output layer is generated by incorporating the curve's sensitivity, specificity, and corresponding area [21]. CNN technique can be used for the pooling parameters. The least-square equation is used for parameter tuning to calculate the case residual and sum it up to get the model residual. Then use the solver add-in to estimate the new values for parameters. The F1 score is also used to check the balance in the classification [22].

E. Final integration model

HashMap containing the mappings of psychometric and social media predictions from the training phase and the forecasts from data collected in real-time from users are used as inputs. The assignment of weights to the response and necessary transformations are done. An artificial neural network (ANN) technique can be used for classification, and dense layers classify data. The train test split is performed in the ratio of (70:30). The grid search determines initial values for parameters and uses an optimizer to minimize residual values. The model is multi-class; thus, the last layer uses a softmax activation that gives the probability of the illness. The HashMap-based model is then trained and deployed in the cloud so that inputs from the web interface are sent to the cloud to predict the final result.

F. Web Interface

A web page is created to scrape tweets using a twint package based on username, and a web page to fetch and display questionnaire data from the database is developed. As shown in Figure 1, if data is present from the psychometric test and Twitter/Reddit, the text data is sent to the cluster ensemble model deployed in the cloud, and psychometric quiz answers are sent to the CNN in the cloud. The results from both models are compared. Then, the weights are assigned to them based on their integrity. Finally, the input is sent to the ANN model in the cloud. Finally, the results are fetched from the cloud, and the probability of getting the mental illness is displayed in the front-end.

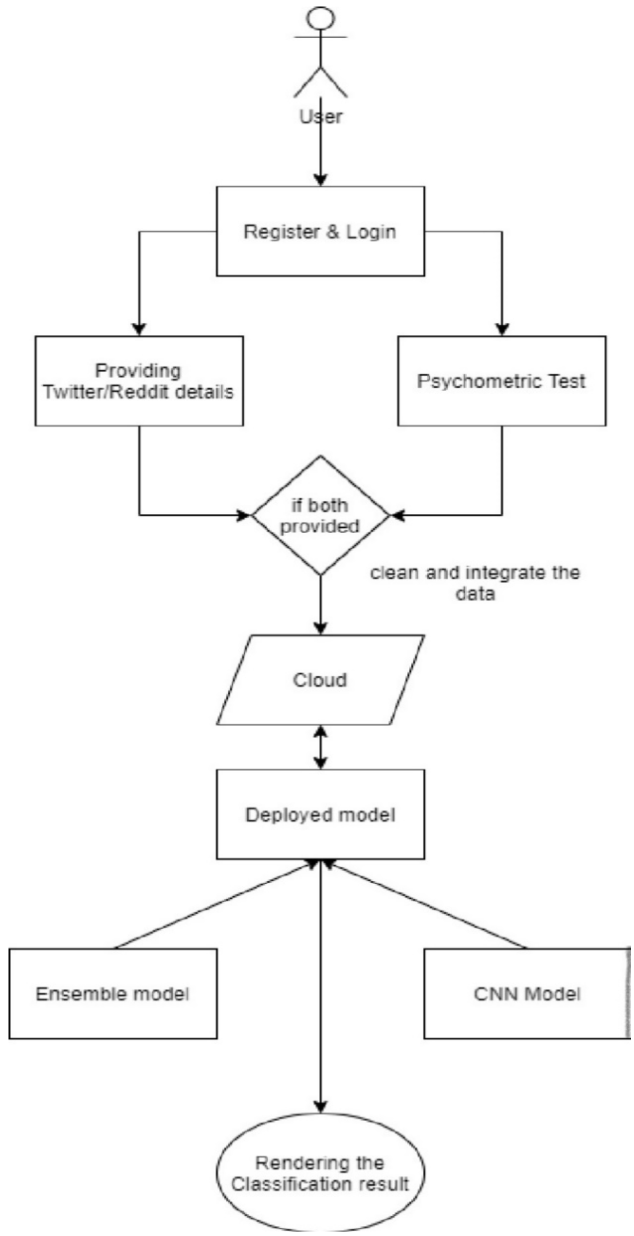

Fig. 1 Workflow diagram of the proposed method

III. RESULTS AND DISCUSSION

A. Dataset Description

- Depression anxiety stress scale responses: The dataset in Kaggle contains psychometric questions and answers submitted by users, and the dataset classifies users based on their mental state.
- Myers-Briggs's personality type dataset: Classifies the user taking the psychometric test into five big traits in which neuroticism is mainly concentrated.
- Sentiment140 dataset with 1.6 million tweets: It contains tweets on which sentiment analysis can be done, and specific keywords relating to each mental illness can be extracted.
- Reddit API: The subreddits posted are extracted by a user and used to find if the user has a mental illness or not.

The whole process is done by implementing two models. First, a cluster ensemble model is being implemented, and it is trained to analyze sentiment and evaluate models using various metrics. The tweets for the past two weeks are scrapped along with the conditions of specific parameters regarding mental illness keywords using the twint package. Some preprocessing is performed for the scrapped tweets, like removing punctuations, extra spaces, stop words, HTML tags, tokenization and stemming, and lemmatization [23]. Converting emojis into text in python can be done using the demoji module. The Emojis are removed and replaced accurately in text strings. Figure 2 shows how the demoji module converts the emoji in the tweet into an equivalent text for further processing.

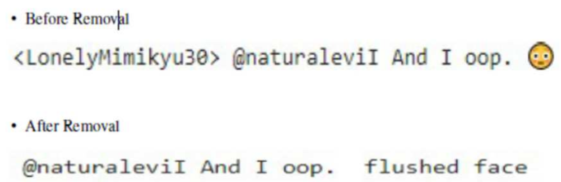

Fig. 2 Emoji to text conversion

Bidirectional encoder representations from Transformers (BERT) are open-source machine learning frameworks for natural language processing [24]. It is designed to help computers understand the meaning of ambiguous language by using surrounding text to establish context. Word2Vec generates the same single vector for the word bank. At the same time, the BERT model generates two different vectors for the word bank. Because of this reason, the BERT model has been implemented. The BERT model divides the text data into 768 columns [25]. Each tweet is converted into an array of vectors with 768 columns, as shown in Figure 3. All the arrays are finally appended inside a higher dimension array. Later, it is passed to t-SNE and principal component analysis [26].

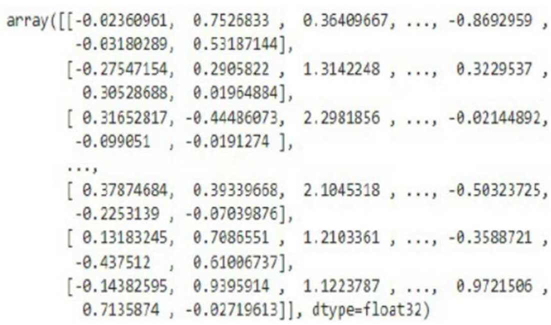

Fig. 3 Vector representation of each tweet generated by the BERT model

K-means computes the centroids and iterates until finding an optimal centroid [27]. The data points are assigned to a cluster so that the sum of the squared distance between the data points and centroid would be minimum [28]. As listed in Table 1, the t-SNE passed k-means has a higher silhouette score than the principal component analysis.

TABLE I
K-MEANS TEST RESULTS

K-means parameter	Silhouette score
t-SNE	0.399
PCA	0.361

In DBSCAN, the silhouette score is negative, and as shown in Figure 4, the final clusters are highly overlapping, and there are more than 217 clusters. OPTICS silhouette score is also negative.

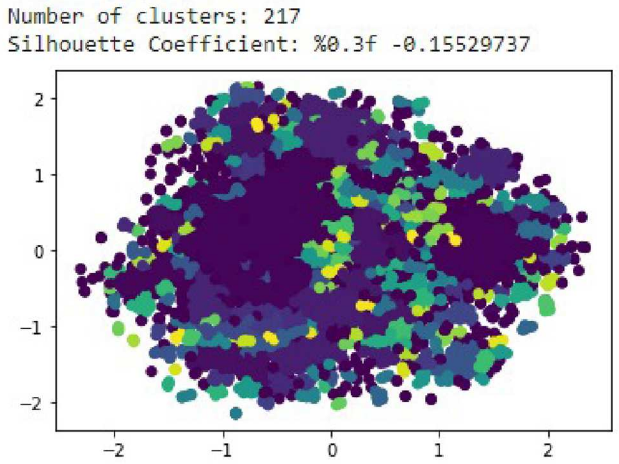

Fig. 4 DBSCAN cluster visualization

Gaussian distribution is the normal distribution, and model-based clustering is an iterative method to fit a set of datasets into clusters [29]. This method works in three steps: First, randomly choose Gaussian parameters and fit them to a set of data points. As listed in Table 2, the silhouette score is positive, and the calinski-harabasaz score is also high in this case.

TABLE II
GAUSSIAN DISTRIBUTION PERFORMANCE SCORE

Silhouette score	Calinski Harabasz score
0.3451	4161.7808

A cluster ensemble package for combining multiple partitions into consolidated clusters is used [30]. The combinatorial optimization problem of obtaining consensus clustering is reformulated in approximation algorithms for the graph or hyper-graph partitioning.

As shown in Figure 5, matrix factorization-based consensus clustering has a higher value than other clustering techniques. The dataset used here is the big five personality test. It has 1000000 rows, each with responses to 50 psychometry survey questions across five different psychometry classes.

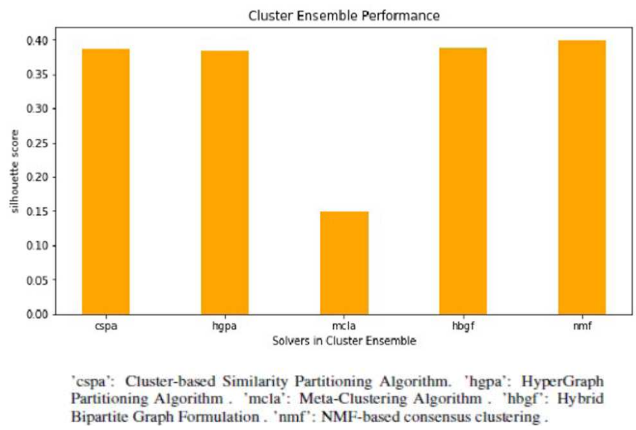

Fig. 5 Silhouette score comparison of different cluster ensemble solvers

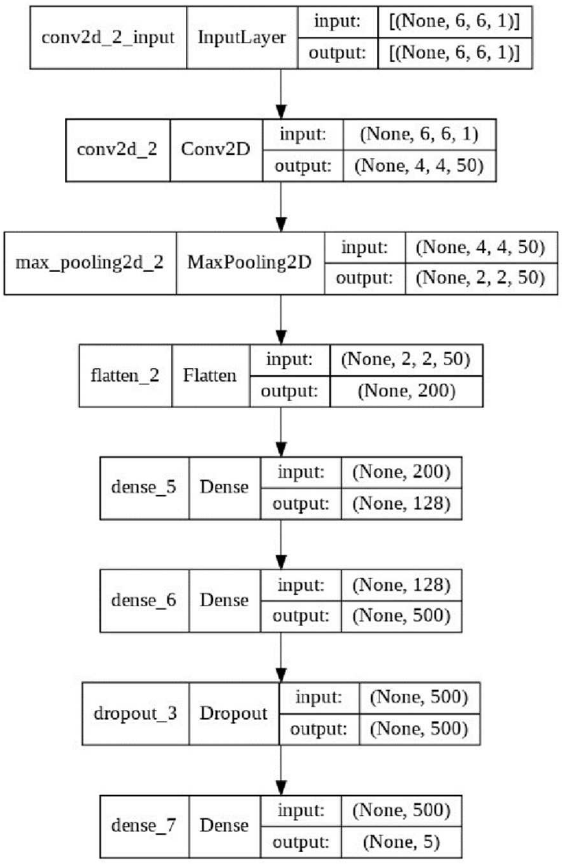

Fig. 6 Architecture of proposed neural network

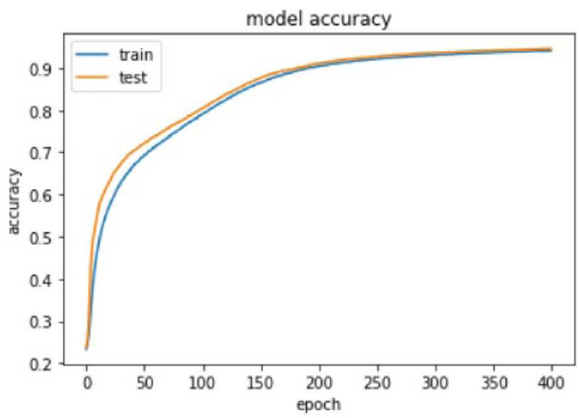

Fig. 7 Accuracy vs Epochs

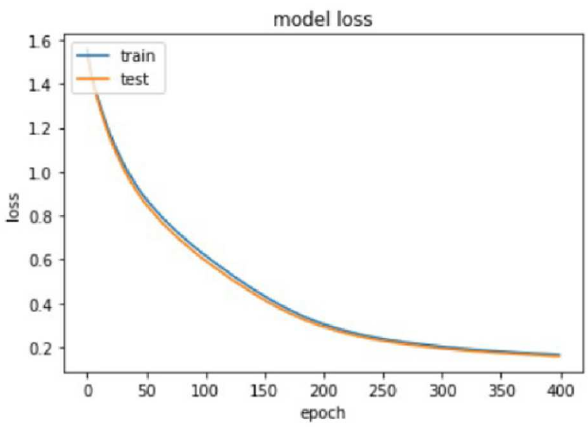

Fig. 8 Loss vs. Epoch

```
EXT1   I am the life of the party.
EXT2   I don't talk a lot.
EXT3   I feel comfortable around people.
EXT4   I keep in the background.
EXT5   I start conversations.
EXT6   I have little to say.
EXT7   I talk to a lot of different people at parties.
EXT8   I don't like to draw attention to myself.
EXT9   I don't mind being the center of attention.
EXT10  I am quiet around strangers.
EST1   I get stressed out easily.
EST2   I am relaxed most of the time.
EST3   I worry about things.
EST4   I seldom feel blue.
EST5   I am easily disturbed.
EST6   I get upset easily.
EST7   I change my mood a lot.
EST8   I have frequent mood swings.
EST9   I get irritated easily.
EST10  I often feel blue.
AGR1   I feel little concern for others.
AGR2   I am interested in people.
AGR3   I insult people.
```

Fig. 9 Questionnaire attributes in psychometry data

Simultaneously a CNN model has been developed and trained to classify data into five main behavior traits and evaluate using an F score.

TABLE III
THE EIGENVALUE FOR DIFFERENT NUMBER OF FACTORS

Factor	**Eigenvalue**
1	5.73289
2	3.65127
3	2.36268
4	1.68417
5	1.41671
6	0.46748
7	0.09957
8	0.05578
9	0.00891
10	-0.04340
11	-0.08229
12	-0.10573
13	-0.12609
14	-0.16681
15	-0.17850
16	-0.19967
17	-0.23917
18	0.26680
19	-0.26816
20	-0.28796

B. Factor Analysis

It is a type of dimensionality reduction. The technique extracts the maximum common variance from all the variables and puts them into a common score. The KMO and Bartlett test evaluate all available data together. A KMO value over 0.5, as shown in Table 4, and a significance level for Bartlett's test below 0.05, as shown in Table 5, suggests a substantial correlation in the data. In the test done with a tool called jamovi, the restriction measures like KMO and Bartlett have been found to pass the conditions mentioned above. The tool gives a pictorial representation of the results. Then to see the number of factors, the eigenvalue is found as listed in Table 3.

TABLE IV
KMO TEST FOR ATTRIBUTES

Attributes	MSA
Overall	0.894
EXT1	0.934
EXT2	0.924
EXT3	0.949
EXT4	0.939
EXT5	0.945
EXT6	0.919
EXT7	0.928
EXT8	0.899
EXT9	0.913
EXT10	0.946
EST1	0.906
EST3	0.904
EST5	0.944
EST6	0.918
EST7	0.853
EST8	0.845
EST9	0.914
EST10	0.946
AGR2	0.920
AGR4	0.858
AGR5	0.876
AGR6	0.900
AGR7	0.880
AGR8	0.936
AGR9	0.872
CSN1	0.883
CSN2	0.836
CSN4	0.916
CSN5	0.882
CSN6	0.849
CSN7	0.858
CSN9	0.847
OPN1	0.700
OPN5	0.787
OPN7	0.885
OPN8	0.693
OPN10	0.784

The CNN is used after the dataset is preprocessed, as shown in Figure 6. Initially, the dataset is divided into a training set and validation set with data aligned training set. Further, the dataset is reduced to a matrix format to comply with the dataset. The input layer is the Conv2D layer for the psychometry classification model with input shape (6, 6, 1). Then, this is sent to a convolution layer with 50 filters of kernel size (3, 3) with a corresponding max-pooling layer. The result is flattened and sent to two sets of dense layers with ReLU activation and then sent to the output layer with sigmoid activation, classifying data into five classes. The performance results are shown in Figure 7 and Figure 8.

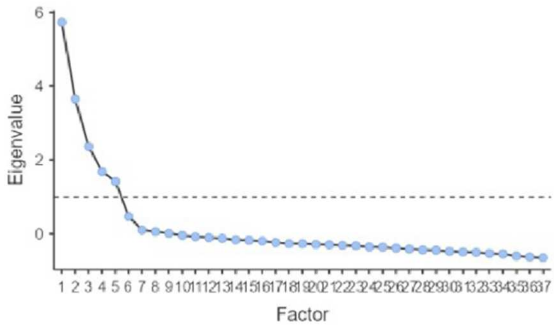

Fig. 10 Factor analysis

The questionnaire attributes in psychometry data are shown in Figure 9. The eigenvalue measures how much of the variance of the observed variables a factor explains. Hence, the factors with an eigenvalue 1 explain more variance than a single observed variable. Those variables for which the factors exceed the value are considered, as shown in Figure 10. The factors that explain the least amount of variance are generally discarded.

TABLE V
BARTLETT'S TEST OF SPHERICITY

X^2	df	p
1.46e+7	666	<.001

IV. CONCLUSION

The remedy suggested in this article builds on the existing system of human observation by incorporating an ANN-powered detection system. The web-based classifier's extensive reach and accessibility to the general public would guarantee that mental illness is more frequently recognized. It would encourage the user to take medication before their condition gets worse. Given that these symptoms are not immediately apparent and may worsen over time, it may also serve as a tool for diagnosing mental illness and encouraging users to seek additional therapy. Only specific users have access to text data from social media platforms. Image and video data from various social media sites can be employed to examine attitudes that would be used to forecast behaviour.

REFERENCES

[1] R. Thorstad, and P. Wolff, "Predicting future mental illness from social media: A big-data approach," Behavior research methods, vol. 51, no. 4, pp. 1586–1600, 2019.

[2] J. O. Pinzón-Arenas, R. Jiménez-Moreno, A. Rubiano, "Path planning for robotic training in virtual environments using deep learning," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 12, no. 1, pp. 8-15, 2022.

[3] Y. H. Yan, T.W. Chien, Y. T. Yeh, W. Chou, S. C.Hsing, et al., "An app for classifying personal mental illness at the workplace using fit statistics and convolutional neural networks: survey-based quantitative study," *JMIR mHealth and uHealth*, vol. 8, no. 7, e17857, 2020.

[4] Y. L. Lee, W. Chou, T. W. Chien, P. H. Chou, Y. T. Yeh, and H. F. Lee, "An app developed for detecting nurse burnouts using the convolutional neural networks in microsoft excel: population-based questionnaire study," *JMIR medical informatics*, vol. 8, no. 5, 2020.

[5] A. Rajeshkumar, and S. Mathi, "Smart solution for reducing COVID-19 risk using internet of things," *Indonesian Journal of Electrical Engineering and Computer Science*, pp. 474-480, 2022.

[6] I. Storey, "Introductory Analysis of the Rasch Model. In Manage Your Own Learning Analytics," *Springer*, Cham, pp. 41-72, 2022.

[7] A. Khatri, and M. Senthilkumar, “Investigation of home agent load balancing, failure detection and recovery in IPv6 network-based mobility,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7 no. 2, pp. 632-641, 2017.

[8] H. Santamaria-Garcia, S. Baez, C. Gomez, O. Rodriguez-Villagra, D. Huepe, M. Portela, P. Reyes, J. Klahr, D. Matallana, and A. Ibanez, “The role of social cognition skills and social determinants of health in predicting symptoms of mental illness,” *Translational psychiatry*, vol. 10, no. 1, pp. 1–13, 2020.

[9] V. Prajapati, R. Guha, and A. Routray, “Multimodal prediction of trait emotional intelligence–through affective changes measured using non-contact based physiological measures,” *Plosone*, vol. 16, no. 7, e0254335, 2021.

[10] B. Ganesh, B. Bala, et al., “An innovative hearing-impaired assistant with sound-localization and speech-to-text application,” *International Journal of Medical Engg and Informatics*, vol. 14.1, pp. 63-73., 2022.

[11] C. Demiroglu, A. Be. Sirli, Y. Ozkanca, and S. Celik, “Depression-level assessment from multi-lingual conversational speech data using acoustic and text features,” *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2020(1), pp. 1–17, 2020.

[12] M. Gaur, V. Aribandi, A. Alambo, U. Kursuncu, K. Thirunarayan, J. Beich, J. Pathak, and A. Sheth, “Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-SSRS,” *PloS one*, vol. 16 no. 5, e0250448, 2021.

[13] B. Silveira, H. S. Silva, F. Murai, and A. P. C. da Silva, “Predicting user emotional tone in mental disorder online communities,” *Future Generation Computer Systems*, vol. 125, pp. 641–651, 2021.

[14] M. Kamal, S. U. Rehman khan, S. Hussain, S. Tariq, and M. F. Ullah, “Predicting mental illness using social media posts and comments,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 607-613, 2020.

[15] K. C´ osic´, S. Popovic´, M. Šarlija, I. Kesedžic´, and T. Jovanovic, “Artificial intelligence in prediction of mental health disorders induced by the covid-19 pandemic among health care workers,” *Croatian Medical Journal*, vol. 61, no. 3, pp. 279, 2020.

[16] G. Chandrasekaran, T. N. Nguyen, and D. J. Hemanth, “Multimodal sentimental analysis for social media applications: A comprehensive review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, e1415, 2021.

[17] Park, Sung Jin, et al. “Cancer and severe mental illness in low-and middle-income countries: The challenges and outlook for the future.” Psycho-Oncology, vol. 30.12, pp. 2002-2011, 2021.

[18] H. T. Nguyen, and M. Le Nguyen, “An ensemble method with sentiment features and clustering support,” *Neurocomputing*, vol. 370, pp. 155-165, 2019.

[19] M. Juez-Gil, A. Arnaiz-González, J. J. Rodríguez, and C. García-Osorio, “Experimental evaluation of ensemble classifiers for imbalance in Big Data,” *Applied soft computing*, vol. 108, 2021.

[20] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, pp. 1-16, 2021.

[21] J. Jacinth Jennifer, and S. Saravanan, “Artificial neural network and sensitivity analysis in the landslide susceptibility mapping of Idukki district, India,” *Geocarto International*, pp. 1-23, 2021.

[22] Chung, Jetli, and Jason Teo. “Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges,” *Applied Computational Intelligence and Soft Computing*, 2022.

[23] P. N. Anjana, and M. Narayanamoorthi, “Secured Natural Language Processing for Conversion of Unstructured Text into Structured Intelligence,” *Second International Conference on Electronics and Sustainable Communication Systems*, pp. 957-962, 2021.

[24] J. G. Nair, Anu Veena, and Aadithya Vinayak. “Comparative study of Twitter Sentiment On COVID-19 Tweets,” *5th International Conference on Computing Methodologies and Communication*, 2021.

[25] Hegde, M. Vinayak, Shilpa, and M. S. Pallavi, “Extracting Attributes of Students Mental Health, Behavior, Attendance and Performance in Academics During COVID-19 Pandemic using PCA Technique,” *ICT Systems and Sustainability, Springer*, Singapore, pp. 551-561, 2022.

[26] C. Yi, S. Tuo, S. Tu, and W. Zhang, “Improved fuzzy C-means clustering algorithm based on t-SNE for terahertz spectral recognition,” *Infrared Physics & Technology*, vol. 117, 103856, 2021.

[27] R. K. Ojha, S. Srivastava, M. Goyal, L. Kumar, A. Kumar, and C. Prasad, “Improving K-Means Effectiveness and Efficiency with Initialization Estimates of Cluster Centroids,” *2nd International Conference on Smart Electronics and Commn.*, pp. 1086-1091, 2021.

[28] Z. Wu, T. Song, and Y. Zhang, “Quantum k-means algorithm based on Manhattan distance,” *Quantum Information Processing*, vol. 21, no. 1, pp. 1-10, 2022.

[29] L. A. García-Escudero, A. Mayo-Iscar, and M. Riani, “Constrained parsimonious model-based clustering,” *Statistics and Computing*, vol. 32, no. 1, pp. 1-15, 2022.

[30] J. Jia, M. Luo, S. Ma, L. Wang, and Y. Liu, “Consensus Clustering-Based Automatic Distribution Matching for Cross-Domain Image Steganalysis,” *IEEE Transactions on Knowledge and Data Engineering*, 2022.