Engineered Science
DOI: https://dx.doi.org/10.30919/es1604

Ensemble-Based Multimodal Analysis for Depression Detection

Ganesh D. Jadhav,[1,*] Sachin D. Babar[2,*] and Parikshit N. Mahalle[3, 4,*]

Abstract

Depression is a widespread mental health disorder that affects individuals globally; it can be detected using traditional methods as well as advanced technological approaches. The study works on various modalities and methods, like textual, audio, and visual modalities, as well as advanced techniques. This study includes research methodology for depression detection using a multimodal dataset containing textual, audio, and visual modalities. The study also consists of clustering and classifying audio and visual modality features. The proposed system incorporates the Ensemble model, which consists of logistic regression, support vector classifier, random forest, and gradient boosting (LSRG). The same clustering and classification process applies to six audio and two visual features. The study determined the final depression level by averaging the results of all modalities using the Late Fusion method. The multimodal system predicts depression levels using heterogeneous data sources. The Mamdani fuzzy gives 93.10% accuracy on the textual modality for text data of Extended Distress Analysis Interview Corpus (E-DAIC), the ensemble model gives 98.21% accuracy on labeled E-DAIC dataset, and late fusion gives 99.54% accuracy for E-DAIC's PHQ-8 dataset. The study aims to integrate multiple models using better-performing techniques. Also, the study offers important suggestions to patients for using a multimodal depression detection system.

Keywords: Hybrid technique; Mental health screening; K-Means; Late fusion; Health informatics.
Received: 05 March 2025; Revised: 10 May 2025; Accepted: 06 June 2025.
Article type: Research article.

1. Introduction

Depression affects people all over the world. Depression has become a significant global mental health challenge, impacting more than 350 million individuals across the world.[1] There are multiple factors in cases of depression, like a business imbalance, some health issues, and some personal life problems from a life partner, so it requires a supplementary monitoring environment to monitor the mental health state.[2] There are multiple methods to diagnose patients, both traditional and modality-based, for self-reporting. Patients can use that method as a supplementary part to monitor them.[3] Modality-based techniques work on the patient and historical data by applying multiple methods to perform mental health evaluations.[4] The important approach is the multimodal approach, which integrates various models or integrated multimodal datasets like textual, audio, and visual data to diagnose depression.[5] The text-based diagnosis depends on the sentiment analysis based on evaluating social media, posts, or transcribed interviews, which has been used to detect depression.[6] The textual data work only on sentimental scores; it cannot cover all features of the emotional and behavioral behavior of the patient.[7] One thing is that features of emotional expressions like facial gestures and vocal tone cannot easily be detected through text.[8,9] Artificial intelligence (AI)-based extraction of audio and visual features for mental health assessment.[10] Therefore, the study requires adding additional modalities like audio and visual data modalities for the detection of depression by capturing these non-verbal feature indicators of depression.[11-13]

The study focused on a multimodal approach to depression detection. The multimodality data uses text modality on the transcript.csv file, audio modalities, six audio features.csv files, and two video features.csv files. The study initiative's objective is to design and develop a highly accurate method that can diagnose depression accurately by using more modalities and features. This study presents a methodology for clustering audio and visual modality feature data and

[1] *Department of Computer Engineering, (SKNCOE- Vadgaon), SPPU, Pune, 411041, India*
[2] *Sinhgad Institute of Technology, Lonavala, (SPPU-Pune), Pune, 410041, India*
[3] *Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune, 411037, India*
[4] *Dean Academics, Vishwakarma Institute of Technology, Pune, 411037, India*
[*] *Email:* jadhavganesh874@gmail.com (G. D. Jadhav); sdbabar@gmail.com (S. D. Babar); parikshit.mahalle@vit.edu (P. N. Mahalle)

classifying using a hybrid ensemble which consists of logistic regression, support vector classifier, random forest, and gradient boosting models. Additionally, combining the results of all models with the late fusion method. The methodology begins with data loading and Preparation, followed by K-means clustering to segment unlabeled data into distinct groups. Feature extraction and the division of train and test datasets are essential in preparing data for machine learning model training. A voting-based ensemble classifier is constructed using multiple models, including logistic regression (LR), support vector classifier (SVC), random forest (RF), and gradient boosting (GB) classifier, which is employed to predict depression conditional levels. The effectiveness of the models is assessed through performance metrics, including accuracy, precision, recall, and F1 score.[2,14] Furthermore, this study introduces a novel depression level analysis by interpreting clusters as depression levels and computing weighted averages based on cluster probabilities. The same process is applied to six audio and two visual features, with the results averaged across modalities to calculate a final depression score. This multimodal approach seeks to enhance originality to support the expanding body of research focused on improving depression detection accuracy by integrating diverse data sources.[15]

1.1 Speech and audio-based depression detection

The study proposes a detailed review of the use of speech as well as audio modality to find out about depression and also study suicidal cases of depression. The study focuses on the audio-modal characteristics such as pitch, energy, and spectral features, which can reveal emotional states connected to depression. The study elaborates on the speech-based models that work on the subtle vocal cues related to patients' mental health. The study also offers non-invasive real-time depression detection. This study has been foundational in establishing audio data as a critical modality for depression detection, noting that features derived from speech can be reliable indicators of depressive symptoms when analyzed effectively.[8] The study works on speech data and the prediction of depression using a convolutional neural network (CNN). The study evaluates the model of the RECOLA database.[16]

The study uses audio and visual feature expression to develop a dataset for audiovisual emotion recognition associated with depression detection. The data collection and final result detection are based on multiple modalities, such as speech and facial expression, to determine depression. The findings of the multimodal depression detection approach integrate the audio and video data, which provides a better understanding of depression. The study analyzes expression and vocal tones individually to convey emotional distress.[11]

1.2 Textual depression detection

The study focused on text-based analysis to detect depression using social media content like Twitter and Facebook. This study is doing work on social media posts to predict depression. Their study analyzed language patterns, including emotional expression, engagement levels, and linguistic markers, to detect depressive behavior. Their work emphasized the feasibility of using large-scale, publicly available textual data to identify individuals at risk for depression, thus providing a scalable solution to mental health screening.[17] The study is based on textual data from Twitter, which helps determine mental illness and the patient's depressive condition. The study analyzes the tweet's content and tone to use social media activity as a predictive tool for mental health issues. Their work further supports text-based analysis in identifying early signs of depression and monitoring its progression over time.[7]

1.3 Depression detection using machine learning

The work included a review of the machine learning algorithms to determine the depression detection methods. The study review reveals that most of the work uses multimodal data. The work is a detailed review of 'machine learning algorithms' in mental health to determine the key elements and applications for depression detection. This research emphasizes the significance of machine learning algorithms, including support vector machines (SVM), random forests, linear regression (LR), and neural networks, in enhancing the accuracy of depression detection systems. The study also describes combining multiple data modalities to improve prediction accuracy, emphasizing the importance of multimodal approaches in future research.[17] The study predicts depression in facial expressions using machine learning algorithms; the study also provides a representation of age, gender, and ethical information about the dataset.[18] This study proposes a multimodal emotion recognition system utilizing deep neural networks. The research primarily focuses on analyzing emotional states and detecting symptoms of depression using audio and visual data. Initially, we must extract the features using the feature extraction method. The training data contains targets and features and is used to train the model. The study focused on a deep neural network to predict accurate data for mental health assessments.[15]

1.4 Multimodal approaches and depression detection

The study proposed combining artificial intelligence with a multimodal system for mental health depression detection. This system applies to combining audio and visual modalities with artificial intelligence.[23] The work uses the combination of multiple models (audio and visual) to predict the accuracy of depression detection. They also show that it produces the correct results in the assessment of mental health, capturing verbal and nonverbal data to indicate depression.[5]

1.5 Software tools for machine learning in depression detection

This study concentrates on the practical application of a machine-learning model for detecting depression. The study fully focuses on the practical implementation of the model using the Scikit-learn library, which is widely used for machine learning in Python. The study is focused on the detailed design and development of the Scikit-Learn library and its applications in various fields, including mental health research. The study also included the Scikit-Learn library and proposed a more efficient algorithm for classification, regression, and clustering for accurate prediction of the mental state of depression. This study also makes an essential tool for researchers to develop depression detection models.[14] The study also produced a powerful tool for detecting depression.[19]

1.6 Societal and media impact on depression stigma

The study is focused on the achievement of technology's effectiveness on social life and how its representations of depression influence public perception. The study is fully focused on societal and media impact on depression stigma, and also how it positively and negatively affects mental health conditions. Also, social media data is useful to predict the human emotion and mental state of depression. The study also highlights the complex relationship between media content and public understanding of mental health.[13] The study gives information about artificial intelligence, specific health issues, and how AI can be a supplementary part of the medical field. It also offers suggestions for ethical implications related to AI technology.[20]

1.7 Global perspective on the depression

According to a report by the World Health Organization (WHO) (2023) depression is a widespread condition affecting individuals globally. They focused on multiple key facts about depression, noting that it is a common mental disorder, affects 5% of adults, impacts more women than men, and contributes to a higher number of suicide cases. The more effective early detection and intervention tools further motivate and explore multimodal machine learning methods in depression detection. The related work covers the various contributions of modalities like text, audio, and visual data for detecting depression and software tools, as well as the global aspect of depression.

The study focused on the multimodality of depression data. The first section introduced depression health issues and also techniques for the diagnosis of depression. It also introduced the study work of research, including various contributions, including text, audio, and visual data modalities; it also covered software tools, social media, and the global issue of depression. The proposed work section covered data set information and proposed methodology for depression detection. The results and discussion section presents all findings from the study, including evaluating each model using performance metrics such as accuracy, precision, recall, F1 score, mean squared error, support, and the average depression level for each model. The suggestion from the study section added multiple suggestions to patients in their lives, and the last section covers the conclusion and future work of the study.

2. Proposed work

The clustering and classification of depression using multimodal features consists of the structure and description of the dataset applied to the system. The proposed methodology contains the clustering and classification using a hybrid ensemble LSRG model. The final result calculated the closest depression level by using the average depression level of all models average the depression level.

2.1 Dataset

The study consists of the clustering and classification of a multimodal feature dataset. The study selected the audio and visual features from the Extended Distress Analysis Interview Corpus (E-DAIC) dataset.[21] The package includes 275 folders of sessions 300–716. Transcript.csv was used in a previous research study.[22] The study selects features consisting of six audio and two visual features. The six audio feature modality contains BoAW_openSMILE_2.3.0_MFCC.csv, BoAW_openSMILE_2.3.0_MFCC.csv, densenet201.csv, OpenSMILE2.3.0_egemaps.csv, OpenSMILE2.3.0_mfcc.csv, vgg16.csv, and two visuals contain BoVW_openFace_2.1.0_Pose_Gaze_AUs.csv and OpenFace2.1.0_Pose_gaze_AUs.csv.[22]

2.2 Proposed methodology

The proposed system architecture for depression detection, illustrated in Fig. 1, is based on an ensemble hybrid model that integrates Logistic regression, Support vector machine, Random forest, and Gradient boosting (LSRG). This ensemble approach enhances performance metrics when applied to the labeled dataset from the Extended Distress Analysis Interview Corpus (E-DAIC).[2] Additionally, the framework incorporates sentiment analysis of textual data using Mamdani fuzzy logic, combined with a CNN and a Bi-directional long short-term memory (Bi-LSTM). The methodology of depression detection using the multimodal dataset works on three modalities: textual, visual, and audio. The work on text modality was explored in the previous work contribution.[22]

2.2.1 Text data modality

The following algorithm explores a process on textual data to get depression detection using sentiment analysis with Mamdani fuzzy logic.[26] The algorithmic process Algorithm 1 extracts the depression detection classification using text data. The above algorithm uses two different methods for sentimental analysis: the first is custom depression detection using a sentimental score threshold, and the second is Mamdani fuzzy logic classification. After the classification of the depression level, the next step is to apply deep learning algorithms like Bi-LSTM and CNN to get results of two learning models. These models will process the classified data to enhance accuracy in predicting depression levels based on

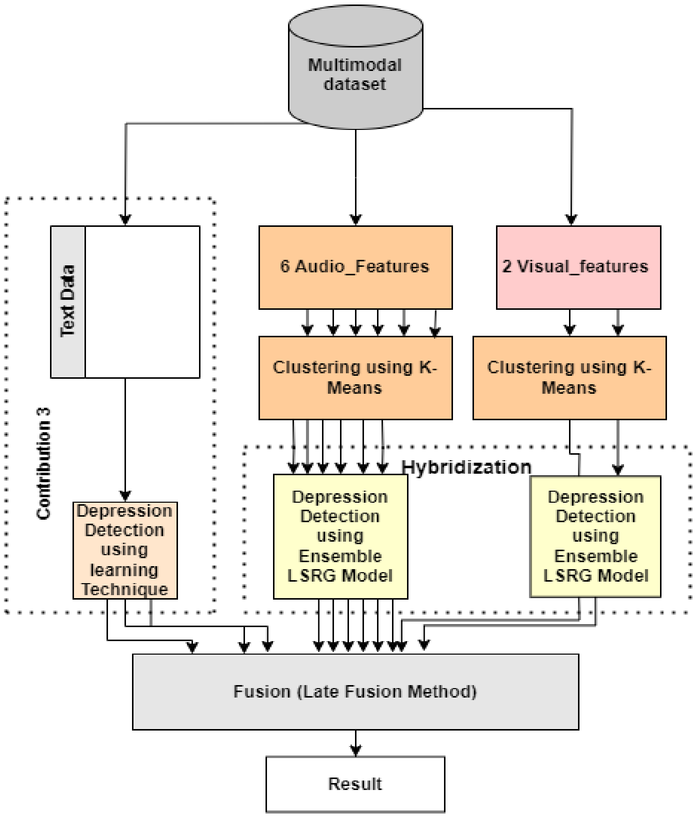

Fig. 1: System architecture.

Algorithm 1: Depression detection using sentiment analysis with Mamdani fuzzy logic algorithm.

Explore process on textual data to get depression detection using sentiment analysis with Mandani fuzzy logic as follows:
Initiate the process of depression detection of text data.
Step 1: Initialization of process: Import all libraries `pandas`, `nltk`, `sentiment.vader`, `matplotlib.pyplot`, `scipy`.
Step 2: Data Loading and Preprocessing: load ⟶ merge_Transcript.csv
Step 3: Sentimental analysis: sentimental analyzer assigns sentimental score (-1 to +1).
Step 4: Mapping Sentimental score to depression levels: Level 3- If Sentimental score>= 0.9; Level 2- if Sentimental score>= 0.6; Level 1- if Sentimental score>= 0.3; Otherwise 0 level.
Step 5: Apply Mamdani fuzzy logic again for classification of depression level. Mamdani fuzzy logic is used to define membership functions for linguistic variables (positive, neutral, negative)
Step 6: Model evaluation and Visualization: Model building and evaluation by applying Bi-LSTM and CNN.
Step 7: Comparison: Comparison of Mamdani fuzzy logic with custom depression detection by using classification_report, accuracy of Bi-LSTM and CNN.
Step 8: End

textual input. By combining the strengths of both deep learning techniques, the study aims to achieve a more robust understanding of sentiment and improve overall detection performance.

2.2.2 Audio-visual data modality

The process involves several stages for clustering and classifying audio and visual modality features.

- Loading and Preparation of data: The code loads the dataset CSV file from the device using file path. After the data loading, the next process is data preparation for clustering and classification. In data preparation, the first two unutilized columns of the dataset are dropped, which is unnecessary for clustering and classification. The remaining columns of the DataFrame are converted into a NumPy array of features for clustering.
- K-Means clustering: The prepared data is unlabeled, so the K-Means clustering algorithm is on the dataset. K-means with 4 clusters applied to the features. Cluster assignments for each data point are calculated. The cluster size is printed, and the largest cluster is found.
- Cluster assignments to the modified dataset: After the cluster assignment, the next process is to save it into the DataFrame. A new column, cluster number, is added to the DataFrame containing K-Means cluster assignments. The modified dataset is optionally saved to a CSV file for future analysis.

- Feature extraction and train-test split data: The study extracts the features (X) by dropping the target variable (y) and cluster number columns. Then, the train_test_split function splits the dataset into training and testing sets. The data is divided into training (80%) and testing (20%) sets. The main goal of this step is to divide the dataset to train the model and evaluate its performance.
- Model initialization and voting classifier: Initialize the four individual classifiers, LR, SVC, RF, and GB classifiers, and define an ensemble voting model. A Voting Classifier is created with soft voting (`voting=` soft`), which averages the predicted probabilities of the classifiers. This step combines different machine learning models, LR, SVC, RF, and GB, into an ensemble LSRG model using soft voting.

The ensemble LSRG model combines the following machine learning models:

a) L - Logistic regression (LR): A LR is a simple and effective linear model of machine learning used for binary and multi-class classification. It predicts probabilities and is well-suited for text-based sentiment and depression level analysis on multimodal data.

b) S - Support vector classifier (SVC): A SVC is a classification algorithm of machine learning techniques that works well with both linear and non-linear data. It finds an optimal hyperplane to separate classes, making it useful for depression level classification on multimodal data.

c) R - Random forest (RF): A RF is an effective ensemble method that generates multiple decision trees and averages their predictions. It is effective for handling structured and unstructured data, including visual and audio features, to find out depression level.

d) G - Gradient boosting (GB): A GB algorithm used to improve weak learners' models sequentially to achieve higher accuracy. It minimizes errors by using multiple iterations, providing an ideal solution for complex depression detection tasks.

The soft voting techniques combine the predicted probabilities of each model instead of just their final predictions. This approach ensures a more balanced and accurate decision, as models contribute proportionally based on their confidence levels.

- Model Training and Prediction: The model trains and predicts. The ensemble LSRG model is trained using the features and target training dataset (*i.e.*, X_train, y_train). The model predicts the clusters for the test dataset (X_test). The functionality of the step is to train the model using training data and predict the clusters for the test data.
- Model Evaluation: In model evaluation, the ensemble LSRG model is evaluated using performance metrics. The performance metrics used in machine learning algorithms for classification of depression level include classification metrics, which consist of accuracy, precision, recall, F1, and error metrics, which consist of Mean Squared Error (MSE). The model evaluation step checks the accuracy and quality of predictions. The study calculates the accuracy using accuracy_score and generates the classification report with classification_report. The classification report shows the precision, recall, and F1 scores. The mean score error is calculated from the difference between predicted and actual clusters.
- Depression Level Analysis: The depression level analysis contains the depression level based on cluster assignments and predicts the average depression level. The count of predictions for each cluster (interpreted as depression levels) is obtained. We calculate the probability of each depression level from 0 to 3 by dividing the count by the total number of predictions. A weighted average depression level is computed based on the probabilities and the predicted depression levels. The closest individual depression level is based on the average depression level.
- Confusion Matrix and Visualization: After calculating the closest depression level, move toward generating and visualizing a confusion matrix to analyze the relationship between true and predicted clusters. The confusion matrix was calculated using confusion_matrix and visualized using seaborn, heatmap for better understanding. The same process follows for six audio and two visual features, and the average depression level and the closest depression level for patients are calculated.

2.2.3 Fusion of text, Audio-visual data modality (Multimodality)

Finally, the depression level was calculated from three textual data models, six audio features, modalities, and two audio feature modalities to compute the final depression level. There is a list of 11 average depression levels created for different modalities. The study calculates the final average depression level by averaging these depression levels using the Late Fusion method. The late fusion method gives better performance metrics than other machine learning and deep learning algorithms.[24,25] Last is the final average depression level compared to the predefined individual levels (*e.g.*, [0, 1, 2, and 3]) to identify the closest depression level.

3. Results and discussion

The work proposed the closest depression level of the patient by applying the late fusion method on textual, audio, and visual modalities. The result of a multimodal analysis often reveals more profound insights by combining various types of data formats, such as text, visual, and audio. The system

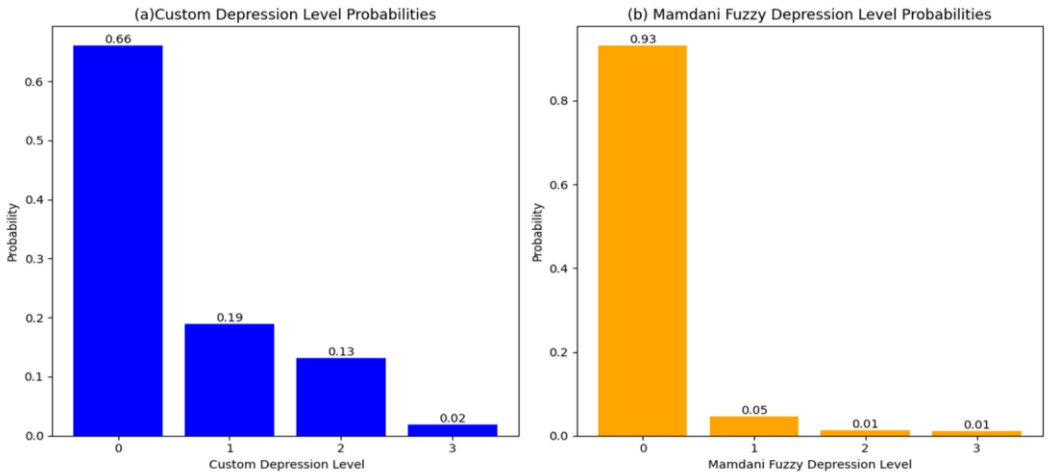

Fig. 2: Probabilities of depression level and Mamdani fuzzy depression level: (a) Custom depression levels and (b) Mamdani fuzzy depression levels.

consists of a proper structural format. The system consists of text modality and audio and visual modalities integrated with the late fusion method. The result of textual modality is described as follows:

3.1 Text data modality

Initially, the discussion was about the result of sentiment analysis. In this system, the sentiment score of each sentence is calculated using the VEDAR sentiment analyzer. The sentiment score is generally between -1 (*i.e.*, very negative) and +1 (*i.e.*, very positive). The study analyzes emotional expression from a textual dataset. The next step is the calculation of custom depression level classification on the basis of sentimental scores, which map with predefined thresholds. The classification of depression is classified into four levels, *i.e.*, level 0 is very negative sentiment, level 1 is neutral sentiment, level 2 is positive sentiment, and level 3 is very positive sentiment.

Fig. 2(a) shows all levels of custom depression level and Fig. 2(b) shows the Mamdani fuzzy logic-based depression levels, along with their associated probability percentages. Fig. 2(a) shows the confidence level of the custom depression classification method for each level of custom depression classification ranging from 0 to 1. It also provides insight into how the custom depression classification technique associates an input text with the specific depression level. Fig. 2(b) shows the output of the probability distribution of each level based on the Mamdani fuzzy depression classification system. This method classified each sentence from the text dataset by using the membership function of linguistic variables.

The membership function for linguistic variable vs. membership degree is shown in Fig. 3. These figures illustrate the varying degrees of membership across different levels of depression, highlighting how effectively the Mamdani fuzzy logic approach can categorize the emotional states represented in the dataset. The membership function defines how input values are mapped to different fuzzy categories, such as no depression, mild, moderate, and severe.

1) X-axis: Represents the numerical input values of sentiment scores, which are related to depression indicators. The sentiment scores are fed into the Mamdani fuzzy logic system.
2) Y-axis: Denotes the degree of membership (ranging from 0

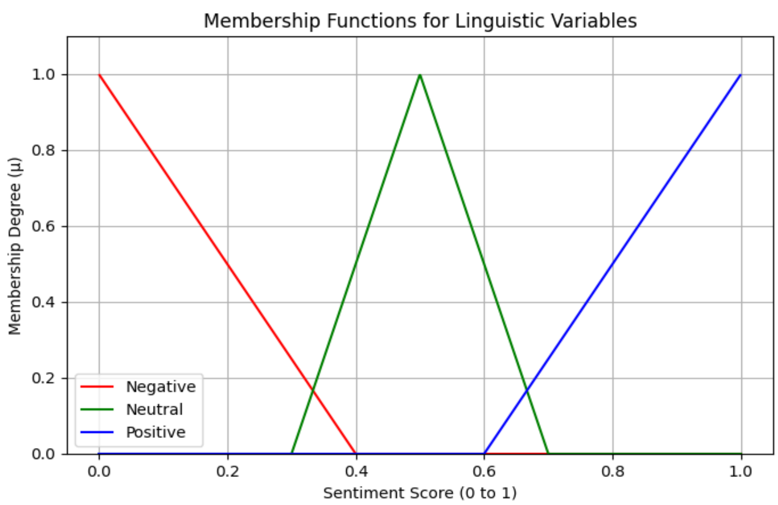

Fig. 3: Membership function for linguistic variable.

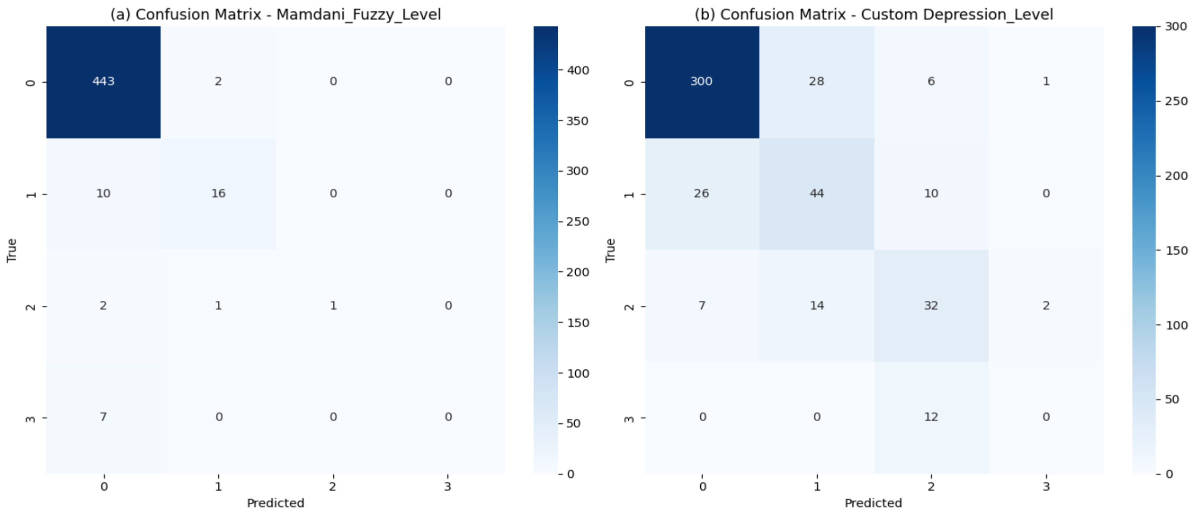

Fig. 4: Confusion matrix of Mamdani fuzzy depression level and custom depression level using CNN: (a) Mamdani fuzzy depression level and (b) Custom depression level.

to 1), which expresses how strongly a given input value belongs to a particular fuzzy category.

Fig. 3 shows each curve on the graph representing the corresponding linguistic term. The triangular shape of the curves defines a piecewise linear transition between linguistic variables. The triangular shape causes the membership degree to increase linearly from zero to one, *i.e.*, full membership, and then decrease linearly back to zero, allowing for a clear yet gradual transition between Mamdani fuzzy classifications. The overlapping of membership functions represents uncertainty in human emotional expression. The smooth transitions allow an input to partially belong to multiple categories, allowing the fuzzy system to handle ambiguity more naturally than custom depression-level systems.

The next step involves exploring both methods using a deep learning algorithm. This study applies Bi-LSTM and CNN algorithms to investigate both approaches and determine which one yields the best performance for depression detection. After applying the CNN to both methods, confusion metrics are generated, as illustrated in Fig. 4. These confusion metrics compare the performance of the two methods: Mamdani_fuzzy_level in Fig. 4(a) and Custom depression level in Fig. 4(b).

Fig. 4 illustrates the performance of CNN on two label sets. Fig. 4 (a) shows that Mamdani fuzzy classification results in a more accurate model with fewer misclassifications, as evident from the higher diagonal values. Fig. 4 (b) shows more confusion across adjacent depression levels, which may stem from the discrete and noisier nature of the custom labels. Mamdani fuzzy depression level classification technique is more effective than the custom depression level classification for CNN.

Fig. 5 shows the accuracy-loss curve of the Mumdani fuzzy level and custom depression level in CNN. In Fig. 5, create a 2 × 2 grid of subplots, *i.e.*, a total of 4 subplots. The CNN model's performance, as illustrated in Fig. 5, shows performance differences between both custom depression level and Mamdani fuzzy depression level classifications. Figs. 5(a) and (c) illustrate that the accuracy of Mamdani fuzzy depression level gives high accuracy and performance values compared to the custom depression Level. Similarly, the loss curves are illustrated in Figs. 5(b) and (d) that the model generates better for the output for the Mamdani fuzzy depression level. These differences can be attributed to the nature of the label's description. The Mamdani fuzzy depression level uses fuzzy logic-based classification, which gives smoother transitions between classes and may reduce ambiguity during training. In contrast, the custom depression level labels are more discrete and possibly noisier, making the classification task slightly more challenging for the CNN. As per the observation, Mamdani fuzzy labeling classification provides better results in feature space, which helps convolutional neural networks capture more discriminative patterns in text data. These observations suggest that fuzzy logic integration can enhance the stability of deep learning models in mental health predictions.

The line plots shown in Fig. 5 represent the training and validation performance of the CNN model across 10 epochs. These curves were derived empirically by recording the model's accuracy and loss at the end of each training epoch for both the training and validation datasets.

Let E be the total number of epochs, and for each epoch $e \in \{1,2,\dots,E\}$, accuracy (A_e) is calculated as in Eq. (1):

$$A_e = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

This metric is computed separately for the training and validation of datasets.

The Loss (L_e) is evaluated using the categorical cross-entropy loss function, defined as Eq. (2):

$$L_e = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{2}$$

where C is the total number of classes, y_i is the true label, and

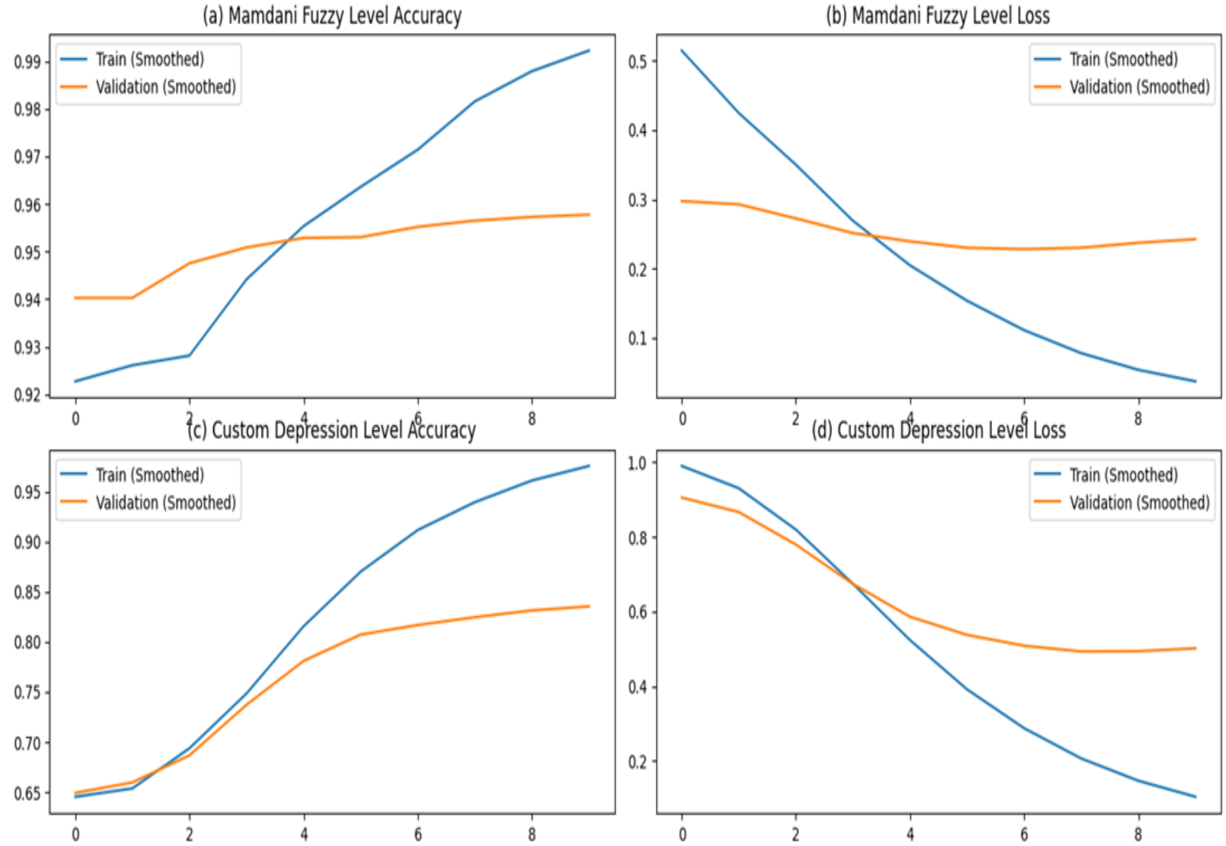

Fig. 5: Accuracy and loss curves of CNN model for both depression level and Mamdani fuzzy depression level classification, (a) Training and validation accuracy for Mamdani fuzzy depression level, (b) Training and validation loss for Mamdani fuzzy depression level, (c) Training and validation accuracy for custom depression level, and (d) Training and validation loss for custom depression level.

$\hat{y}_i$ is the predicted probability for class i (output of softmax layer of CNN).

These values A_e and L_e were computed and recorded after each epoch during the training process from Eqs. (1) and (2) respectively. The resulting data arrays were plotted using Matplotlib to visualize the model's learning progression over time.

Fig. 6 shows the receiver operating characteristic (ROC) curves of both methods in CNN. The ROC plots true positive rate vs. false positive rate at various thresholds. In the ROC plot, a reusable function plots the ROC curve for multiclass classification.

The ROC curve of Mamdani fuzzy depression level classification using CNN in Fig. 6(a) illustrates the model's performance in differentiating between fuzzy depression categories obtained using fuzzy logic principles. The ROC

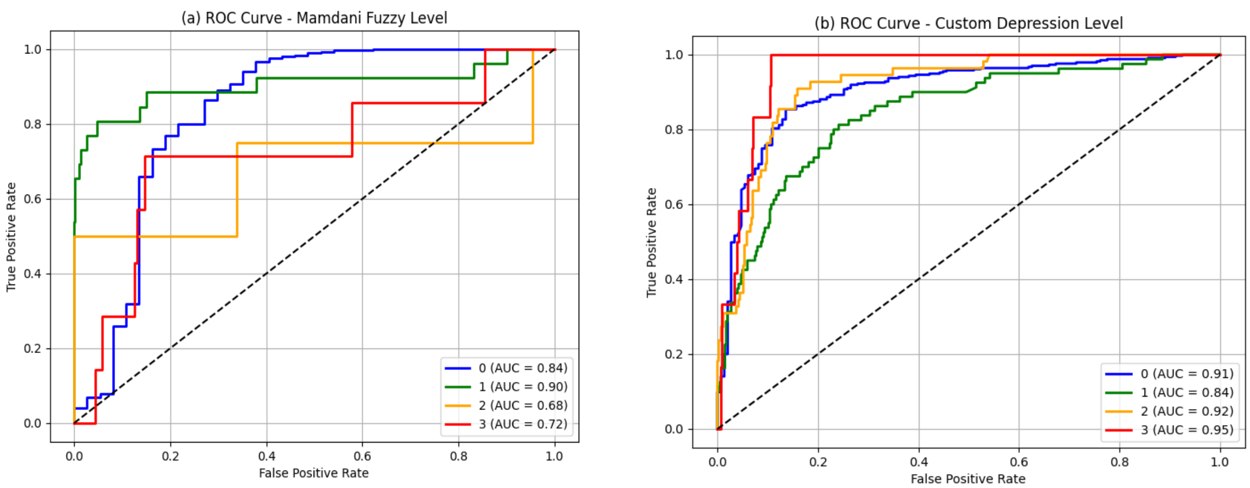

Fig. 6: ROC curve of custom depression level and Mamdani fuzzy depression level using CNN: (a) Mamdani fuzzy depression level and (b) Custom depression level.

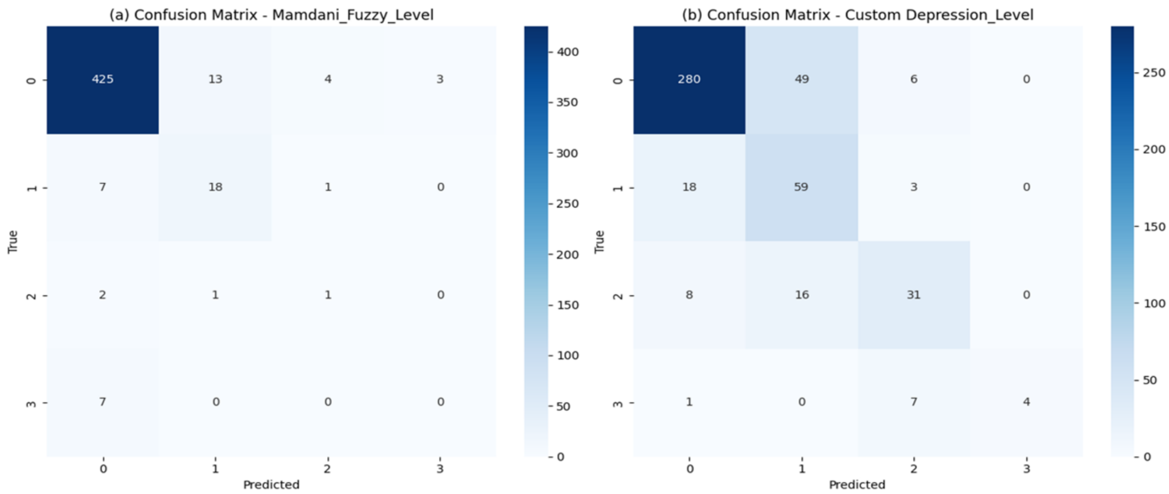

Fig. 7: Confusion matrix of Mamdani fuzzy depression level and depression level using Bi-LSTM: (a) Mamdani fuzzy depression level and (b) Custom depression level.

curve for the Custom Depression level using CNN in Fig. 6(b) represents classification performance based on manually labeled or statistically derived class definitions. This curve may show slightly more overlap between classes compared to the fuzzy model, particularly if the class boundaries are not as clearly defined. The Mamdani fuzzy model benefits more from well-defined boundaries, resulting in higher AUC scores and more distinct ROC curves per class than the custom depression level. The Custom depression model has more class overlap; it has slightly lower AUCs in some categories.

After applying the Bi-LSTM to both methods, confusion metrics are generated, as illustrated in Fig. 7. These confusion metrics compare the performance of the two methods: Mamdani_fuzzy_level and custom depression level. Fig. 8 shows the accuracy-loss curve of the Mumdani fuzzy level and custom depression level in Bi-LSTM. Fig. 9 shows the ROC curves of both methods in Bi-LSTM.

Fig. 7 illustrates the performance of Bi-LSTM on two label sets. Fig. 7(a) shows that Mamdani fuzzy classification results in a more accurate model with fewer misclassifications, as evident from the higher diagonal values. Fig. 7(b) shows more confusion across adjacent depression levels, which may stem from the discrete and noisier nature of the custom labels. Mamdani fuzzy depression level classification technique is more effective than the custom depression level classification for Bi-LSTM.

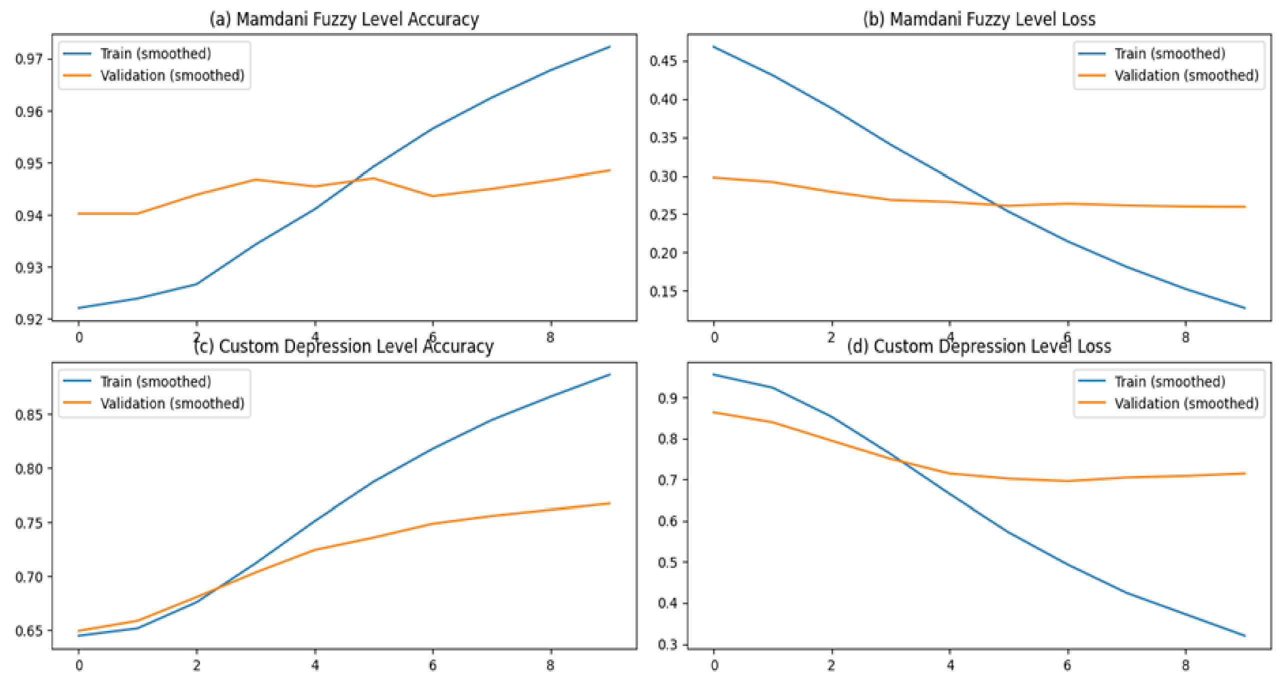

Fig. 8: Accuracy and loss curves of Bi-LSTM model for both depression level and Mamdani fuzzy depression level classification: (a) Training and validation accuracy for Mamdani fuzzy depression level. (b) Training and validation loss for Mamdani fuzzy depression level. (c) Training and validation accuracy for custom depression level. (d) Training and validation loss for custom depression level.

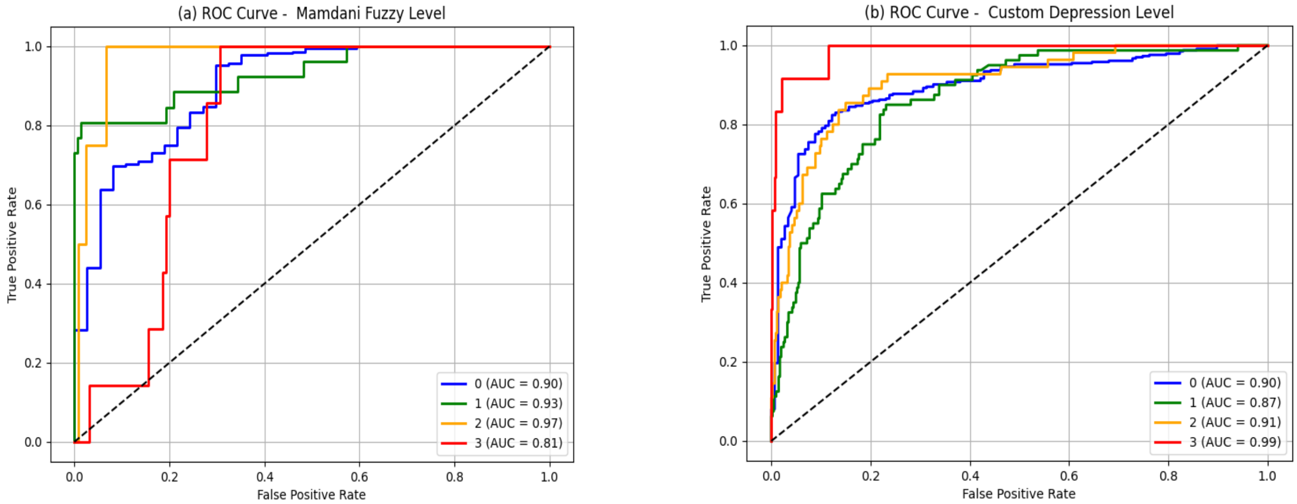

Fig. 9: ROC curve of custom depression level and Mamdani fuzzy depression level using Bi-LSTM: (a) Mamdani fuzzy depression level and (b) Custom depression level.

The Bi-LSTM model's performance, as illustrated in Fig. 8, shows performance differences between both custom depression Level and Mamdani fuzzy depression level classifications. Figs. 8(a) and (c) illustrate that the accuracy of Mamdani fuzzy's depression level gives high accuracy and performance values compared to the custom depression level. Similarly, the loss curves are illustrated in Figs. 8(b) and (d) that the model generates better for the output for the Mamdani fuzzy depression level. The Mamdani fuzzy depression level uses fuzzy logic-based classification, which gives smoother transitions between classes and may reduce ambiguity during training. In contrast, the custom depression level labels are more discrete and possibly noisier, making the classification task slightly more challenging for the Bi-LSTM. As per the observation, Mamdani fuzzy labeling classification provides better results in feature space, which helps convolutional neural networks capture more discriminative patterns in text data. These observations suggest that fuzzy logic integration can enhance the stability of deep learning models in mental health predictions.

The ROC curve of Mamdani fuzzy depression level classification using Bi-LSTM in Fig. 9(a) illustrates the model's performance in differentiating between fuzzy depression categories obtained using fuzzy logic principles. The ROC curve for the Custom Depression level using Bi-LSTM in Fig. 9(b) represents classification performance based on manually labeled or statistically derived class definitions. The Mamdani fuzzy model benefits more from well-defined boundaries, resulting in higher AUC scores and more distinct ROC curves per class than the custom depression level. The custom depression model has more class overlap; it has slightly lower AUCs in some categories.

In textual modality, the performance of custom depression and Mamdani fuzzy logic depression is compared using deep learning algorithms such as CNN and Bi-LSTM, as shown in Table 1. The performance of the custom depression method using CNN and Bi-LSTM is an accuracy of 77.25% and 77.38%, respectively; a precision of 78.00% and 79.00%; a recall of 79.00% and 77.00%; and an F1 score of 79.00% and 78.00% for support value 482. Another performance using Mamdani fuzzy level using CNN and Bi-LSTM is the accuracy of 95.02% and 93.36%, respectively; the precision of 94.00% and 92.00%; recall of 95.00% and 93.00%; and F1 score of 94.00% and 93.00% for support value 482.

The results discussion shows that the Mamdani fuzzy level gives better performance than the custom depression level for both CNN and Bi-LSTM. The current work used only the average depression level of all models like Mamdani fuzzy classification, Convolutional Neural Network with Mamdani fuzzy, and Bi-directional Long Short-Term Memory (Bi-LSTM) with Mamdani fuzzy.

3.2 Audio-visual feature data modality

Depression detection using audio and visual feature data modality consists of a total of eight features; two are visual modality, and six are audio modality. The study takes only one

Table 1: Comparison of performance.

Sr. No.	Methods	Learning techniques	Accuracy in %	Precision in %	Recall in %	F1 score in %	Support
1	Custom depression	CNN	79.25	78.00	79.00	79.00	482
2		Bi-LSTM	77.38	79.00	77.00	78.00	482
3	Mamdani_fuzzy	CNN	95.02	94.00	95.00	94.00	482
4		Bi- LSTM	93.36	92.00	93.00	93.00	482

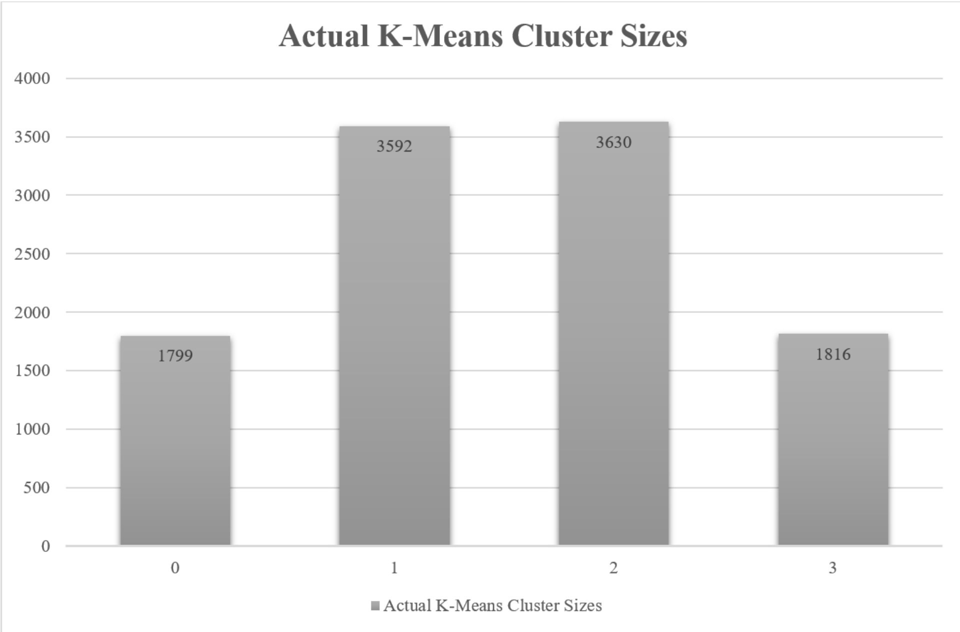

Fig. 10: Number of samples with each cluster in audio modality.

Table 2: Comparison performance of model with existing algorithm (Audio modality).

Sr.	Model	Accuracy in %	Precision in %	Recall in %	F1 Score in %	Mean squared error in %
1	LR	56.36	61.00	56.00	56.00	44.49
2	SVM	97.32	97.00	97.00	97.00	07.84
3	RF	96.90	97.00	97.00	97.00	10.23
4	GB	96.44	96.00	96.00	96.00	10.14
5	CNN	97.41	97.00	97.00	97.00	09.10
6	RNN	97.37	98.00	97.00	98.00	08.48
7	Proposed ensemble LSRG model	97.78	98.00	98.00	98.00	07.24

sample feature to compare the existing algorithm with the proposed method. The study applies clustering and classification algorithms separately to each feature. The system takes data from only one patient and processes each feature by applying k-means clustering and a classification algorithm for result purposes. Initially, for the audio feature modality, take one sample audio feature dataset, *i.e.*, OpenSMILE2.3.0_egemaps.csv, and by applying k-means clustering, the depression level is classified into four clusters as shown in Fig. 10. Generally, it converts unlabeled data to labelled data by applying a clustering algorithm.

After data labelling the cluster system, update the dataset with the cluster number and apply the proposed ensemble LSRG model to predict depression levels accurately. Then, the study compares existing algorithms' performance metrics, such as accuracy, precision, recall, F-1 score, and mean squared error.

Table 2 shows the comparative analysis and performance of the model on one sample audio dataset using an ensemble model and machine learning algorithms. The proposed ensemble LSRG model gives high performance with minimum error as compared to existing machine learning and deep learning algorithms. The proposed model gives 97.78% accuracy, 98.00% precision, 98.00% recall, 98.00% F1 score, and 7.24% mean squared error. Hence, the ensemble model is used for the next process of audio feature data modality.

3.3 Visual data modality

Next step for the visual feature modality, take one sample visual feature dataset, *i.e.*, BoVW_openFace_2.1.0_Pose_Gaze_AUs.csv, and by applying k-means clustering, the depression level is classified into four clusters as shown in Fig. 11. Generally, it converts unlabeled data to labeled data by applying a clustering algorithm.

Table 3 shows the comparative analysis and performance of the model on one sample visual dataset using an ensemble model and machine learning algorithms. The proposed ensemble LSRG model gives high performance with minimum error as compared to existing machine learning and deep learning algorithms. The proposed model gives 99.95% accuracy, 100.00% precision, 100.00% recall, 100.00% F1

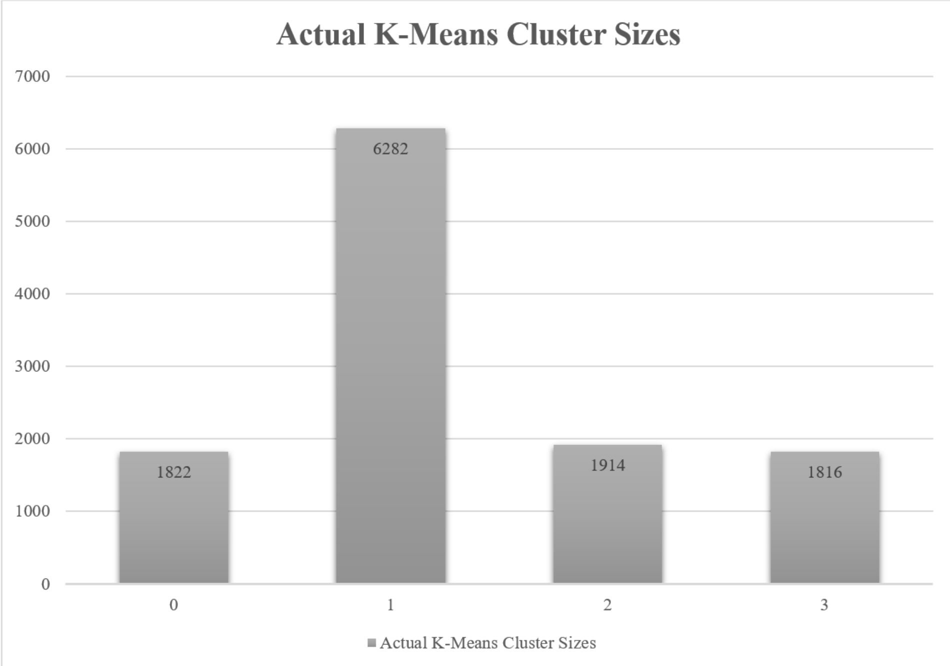

Fig. 11: Number of samples with each cluster in visual modality.

Table 3: Comparison performance of model with existing algorithm (Visual modality).

Sr.	Model	Accuracy in %	Precision in %	Recall in %	F1 Score in %	Mean squared error in %
1	LR	92.38	93.00	92.00	92.00	07.69
2	SVM	99.44	99.00	99.00	99.00	02.29
3	RF	99.86	100.00	100.00	100.00	00.78
4	GB	99.72	100.00	99.00	100.00	01.42
5	CNN	99.72	100.00	100.00	100.00	01.33
6	RNN	99.77	100.00	100.00	100.00	00.91
7	Proposed ensemble LSRG model	99.95	100.00	100.00	100.00	00.41

score, and 0.41% mean squared error. Hence, the ensemble model is used for the next process of visual feature data modality.

3.3 Fusion of audio–visual, text data modality (Multimodality)

The multimodal depression detection system consists of a fusion of multiple modalities, such as textual, audio, and visual, for individual patients.

Textual modality models: The study uses the Mamdani fuzzy level system for multimodal depression detection because it gives high performance and accuracy as compared to custom depression detection. For multimodal detection, three average depression levels from three textual modalities, *i.e.*, average depression level of Mamdani fuzzy logic, average depression level of Mamdani fuzzy logic with convolution neural network, and average depression level of Mamdani fuzzy logic with Bi-LSTM.

Audio and visual models: The work on audio and visual modalities builds the ensemble LSRG hybrid model separately for each feature. Since the dataset contains eight features, the system finds the average depression levels of eight ensemble LSRG models for eight features. Model 1 is audio_feature, model 2 is audio_feature, model 3 is visual_feature, model 4 is audio_feature, model 5 is visual_feature, model 6 is audio_feature, model 7 is visual_feature, and model 8 is audio_feature.

Multimodality: The final average depression level is calculated by applying the late fusion method to all models, *i.e.*, three from textual data, six from audio modalities, and two from visual modalities. The final closed individual depression level of patients is calculated using the final average depression level shown in Fig. S1.

In the discussion part, initially, the study focuses on textual data and calculates the sentiment score for sentences. The custom_depression and Mamdani_fuzzy were used for the classification of sentimental data based on the sentimental Score for the calculation of probability for each level, and

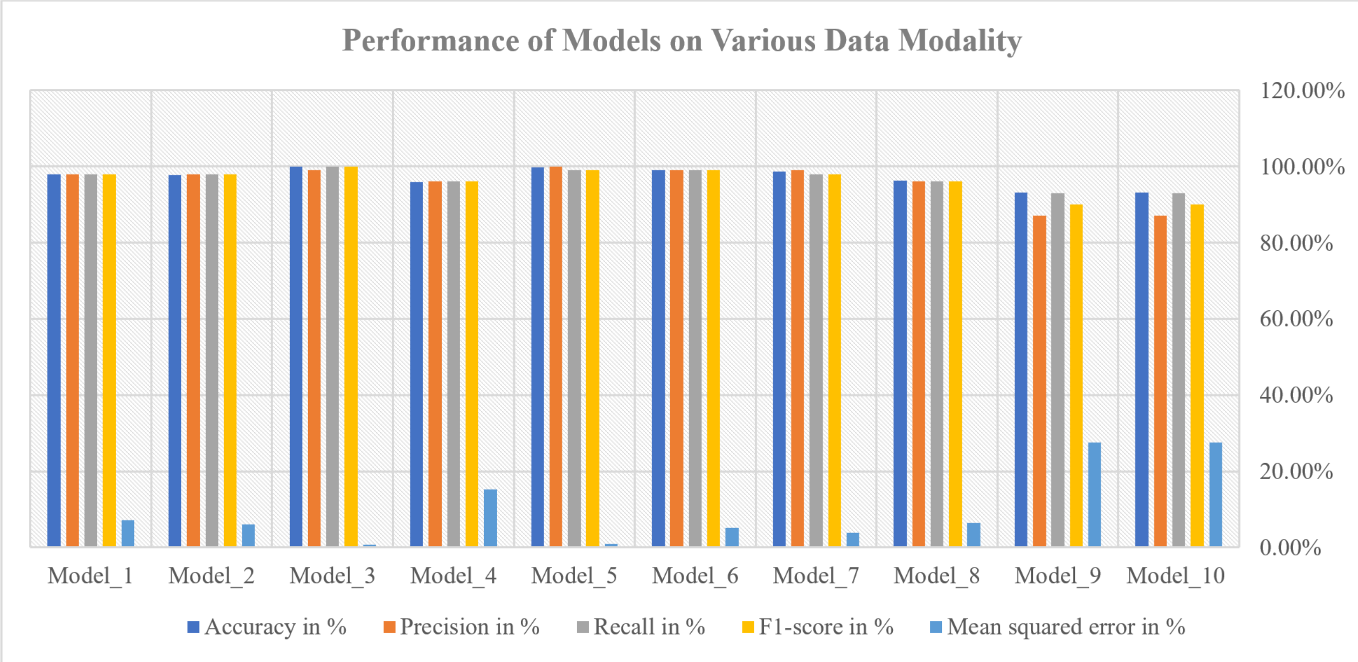

Fig. 12: Performance of models on various data modalities.

Table 4: Performance of models on various all data modalities.

Sr.	Model	Modality	Accuracy in %	Precision in %	Recall in %	F1 Score in %	Mean squared error in %	Support	Average depression level
1	Ensemble Model_1	Audio	97.83	98.00	98.00	98.00	7.19	2168	1.5064
2	Ensemble Model_2	Audio	97.64	98.00	98.00	98.00	6.04	2168	1.6033
3	Ensemble Model_3	Visual	99.86	99.00	100.00	100.00	0.78	2179	1.1835
4	Ensemble Model_4	Audio	95.85	96.00	96.00	96.00	15.20	217	1.5668
5	Ensemble Model_5	Visual	99.75	100.00	99.00	99.00	0.93	6537	1.3910
6	Ensemble Model_6	Audio	99.05	99.00	99.00	99.00	5.21	21785	1.4396
7	Ensemble Model_7	Audio	98.73	99.00	98.00	98.00	3.81	21786	1.3568
8	Ensemble Model_8	Audio	96.31	96.00	96.00	96.00	6.45	217	1.0921
9	Mamdani fuzzy RNN	textual	93.10	87.00	93.00	90.00	27.58	29	0.04
10	Mamdani Fuzzy CNN	textual	93.10	87.00	93.00	90.00	27.58	29	0.04

updated the data with sentimental_analysis_results_with_mamdani_fuzzy.csv. The Mamdani fuzzy gives a better result than the custom depression level. The results are validated using CNN and RNN. The custom_depression with CNN is accuracy, precision, recall, F1_score, and mean_squared_error of 58.62%, 34.00%, 59.00%, 43.00%, and 120.68% respectively. The Mamdai_fuzzy_depression with CNN accuracy, precision, recall, F1_score, and mean_squared_error of 93.10%, 87.00%, 93.00%, 90.00%, and 27.58% respectively. The custom_depression with RNN accuracy, precision, recall, F1_score, mean_squared_error, of 58.62%, 34.00%, 59.00%, 43.00%, and 120.68% respectively. The Mamdai_fuzzy_depression with RNN, accuracy, precision, recall, F1_score, mean_squared error of 93.10%, 87.00%, 93.00%, 90.00%, and 27.58% respectively. Table 1 analyzes that Mamdani fuzzy depression is better than custom_depression with CNN and RNN. The calculation of the average depression level of Mamdani_fuzzy is based on probability for each level of depression.

The sentimental_analysis_results_with_mamdani_fuzzy. csv file is used as input for both the RNN and CNN to calculate the average depression level using Mamdani fuzzy for RNN and CNN, respectively. Fig. 4 and Fig. 7 show the confusion metrics of Mamdani_fuzzy vs. Custom depression with RNN and CNN, respectively. Finally, Fig. S1 shows the calculation of the average depression level of all 11 models from different features and calculates the closest individual depression level for depression.

Table 4 and Fig. 12 show the performance of models with various performance metrics. Model_1 to Model_8 were developed by using a hybrid ensemble LSRG model. According to our previous study,[2] the hybrid LSRG model achieved performance metrics of 98.21% accuracy, 99% precision, 99% recall, 99% F1 score, and 1.78% mean squared error on the labeled dataset from the Extended Distress Analysis Interview Corpus (E-DAIC). The hybrid LSRG model performs better than Linear_regression, Support_vector_classifier, random_forest_classifier, Gradient_boosting classifier, CNN, and RNN on E-DAIC label dataset as per our previous study.[2] The late fusion method gives the performance metrics like accuracy, Precision, Recall, F1_score, mean_squared_error, mean absolute error, R2 score of 99.54%, 99%, 99%, 99%, 0.45%, 0.45%, and 99.46%, respectively and it is better than Linear_regression, Support_vector_classifier, random_forest_classifier, Gradient_boosting classifier, CNN and RNN on Detailed_PHQ8_Labels.csv file of E-DAIC as per experiment.[25]

The challenging task of a multimodal system is calculating the result using heterogeneous data with different support values, as shown in Table 4. The system calculates the final closest depression level by combining the visual and audio modalities and the average depression levels. The study applies Mamdani fuzzy logic, *i.e.*, model_11, Mamdani fuzzy with RNN, *i.e.*, model_9, and Mamdani fuzzy with CNN, *i.e.*, model_10 to the textual modality. The study aims to integrate multiple better-performance models using better-performance techniques.

4. Conclusion

The proposed multimodal system contains a multimodal dataset comprising textual, *i.e.*, transcript.csv, audio, and visual features to improve prediction using a late fusion of the hybrid learning algorithm and the Mamdani fuzzy algorithm with CNN and RNN. The study used diverse data sources, such as textual, audio, and visual. Initially, the process involved textual data using sentiment scores, and Mamdani fuzzy logic was used to determine an individual's mental state of depression. The proposed methodology incorporates audio and visual features, utilizes the K-Means algorithm for data clustering, and updates the data with cluster labels.

An ensemble voting classifier is then applied for model prediction, resulting in good accuracy in detecting depression levels. The audio and visual features are used in an ensemble LSRG model to predict the average depression level accurately. The final result calculates the average depression level by using the result of the hybrid learning techniques for audio + visual features and Mumdani fuzzy logic models on a multimodal dataset in Fig. S1, Table 1, and Fig. 12 show the performance metrics of the different parameters like accuracy, precision, recall, F1 score, mean squared error, support, and the closest depression level. All models give better accuracy for prediction. Future studies should explore the combination of additional modalities, features, and applications to solve real-world problems, such as in clinical settings or through digital mental health interventions.

The study proposed a multimodal depression detection system based on textual, audio, and visual data modality. The system used to detect depression. The advice for patients who may use the multimodal depression system is to treat it as a supplement, not a replacement for professional medical advice. Using a multimodal system, professionals consult with people to detect depression, interpret the result, and prepare a comprehensive treatment plan. This system is helpful for self-monitoring of a patient's mental health. It also identifies the patient's emotional state, supports early detection of depression, and helps patients prepare a treatment strategy. The system's use is consistent, so it works effectively and is useful to provide more accurate results. The system can monitor patients and their daily activities, providing a comfortable and manageable experience.

Depression is a mental state that can fluctuate in a patient when the system gives them too much daily. If patients use any of the systems or any app or platforms to handle their mental data, ensure they guarantee the privacy and security of their personal information. Discuss with the platform provider or clinician so they can give proper guidance on using the platform. This system can help in your treatment plan, and your therapists and doctors can use this data for appropriate treatment or suggest meditation and lifestyle changes. If the patient gets a good experience and is more accurate with the system, they must participate by providing feedback on the results and correcting any misinterpretations. This section offers multiple suggestions to the patients regarding integrating and using the multimodal system to monitor their emotional well-being and maintain professional support.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Applicable.

References

[1] G. Jadhav, S. Babar and P. Mahalle, A survey: Performance-aware depression detection, *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, March 15-17, 2023, 1242-1249. doi:10112572.

[2] G. D. Jadhav, S. D. Babar, P. N. Mahalle, Hybrid approach for enhanced depression detection using learning techniques, *International Journal of Advanced Computer Science and Applications*, 2024, **15**, 892-900, doi: 10.14569/ijacsa.2024.0150492.

[3] P. Robinson, D. Turk, S. Jilka, M. Cella, Measuring attitudes towards mental health using social media: investigating stigma and trivialisation, *Social Psychiatry and Psychiatric Epidemiology*, 2019, **54**, 51-58, doi: 10.1007/s00127-018-1571-5.

[4] A. B. R. Shatte, D. M. Hutchinson, S. J. Teague, Machine learning in mental health: a scoping review of methods and applications, *Psychological Medicine*, 2019, **49**, 1426-1448, doi: 10.1017/S0033291719000151.

[5] X. Xu, J. Li, Z. Zhu, L. Zhao, H. Wang, C. Song, Y. Chen, Q. Zhao, J. Yang, Y. Pei, A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis, *Bioengineering*, 2024, **11**, 219, doi: 10.3390/bioengineering11030219.

[6] F. Abu Bakar, N. Mohd Nawi, A. A. Hezam, Predicting depression using social media posts, *Journal of Soft Computing and Data Mining*, 2021, **2**, 39-48, doi: 10.30880/jscdm.2021.02.02.004.

[7] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, E. J. Langer, Forecasting the onset and course of mental illness with Twitter data, *Scientific Reports*, 2017, **7**, 13006, doi: 10.1038/s41598-017-12961-9.

[8] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T. F. Quatieri, A review of depression and suicide risk assessment using speech analysis, *Speech Communication*, 2015, **71**, 10-49, doi: 10.1016/j.specom.2015.03.004.

[9] A. Atmakuru, A. Shahini, S. Chakraborty, S. Seoni, M. Salvi, A. Hafeez-Baig, S. Rashid, R. S. Tan, P. D. Barua, F. Molinari, U. R. Acharya, Artificial intelligence-based suicide prevention and prediction: a systematic review (2019–2023), *Information Fusion*, 2025, **114**, 102673, doi: 10.1016/j.inffus.2024.102673.

[10] S. Dhelim, L. Chen, H. Ning, C. Nugent, Artificial intelligence for suicide assessment using Audiovisual Cues: a review, *Artificial Intelligence Review*, 2023, **56**, 5591-5618, doi: 10.48550/arXiv.2201.09130.

[11] B. Moghe, M. Kachhara, K. M, Multimodal emotion analysis for depression detection- integrating facial expression and speech recognition, *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, June 11-12, Bangalore, India, IEEE, 2024, 37-42, doi: 10.1109/ICICI62254.2024.00015.

[12] Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, G. Huang, Z. Liang, L. Zhang, L. Li, H. Ding, Z. Zhang, J. Wang, Multimodal sensing for depression risk detection: integrating audio, video, and text data, *Sensors*, 2024, **24**, 3714, doi: 10.3390/s24123714.

[13] H. Wang, Y. Liu, X. Zhen, X. Tu, Depression speech recognition with a three-dimensional convolutional network, *Frontiers in Human Neuroscience*, 2021, **15**, 713823, doi: 10.3389/fnhum.2021.713823.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, É. Duchesnay, Scikit-learn: Machine learning in Python, *The Journal of machine learning research*, 2011, **12**, 2825-2830, doi: 10.48550/arXiv.1201.0490.

[15] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, *IEEE Journal of Selected Topics in Signal Processing*, 2017, **11**, 1301-1309, doi: 10.1109/JSTSP.2017.2764438.

[16] P. Tzirakis, J. Zhang, B. W. Schuller, End-to-end speech emotion recognition using deep neural networks, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 15-20, 2018, Calgary, AB, Canada, IEEE, 2018, 5089-5093, doi: 10.1109/ICASSP.2018.8462677.

[17] S. Biswas, Y. Hasija, Predicting depression through social media, *Predictive Analytics of Psychological Disorders in Healthcare: Data Analytics on Psychological Disorders*, 2022, **128**, 109-127, doi: 10.1007/978-981-19-1724-0_6.

[18] S. Tyagi, K. Panchal, P. Chitta, S. Todi, S. Priya, R. Pawar, Emotionomics: pioneering depression detection through facial expression analytics, *2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, August 18-19, Pune, India, IEEE, 2023, 1-4, doi: 10.1109/ICCUBEA58933.2023.10392157.

[19] E. Sajno, S. Bartolotta, C. Tuena, P. Cipresso, E. Pedroli, G. Riva, Machine learning in biosignals processing for mental health: a narrative review, *Frontiers in Psychology*, 2023, **13**, 1066317, doi: 10.3389/fpsyg.2022.1066317.

[20] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H. C. Kim, D. V. Jeste, Artificial intelligence for mental health and mental illnesses: an overview, *Current Psychiatry Reports*, 2019, **21**, 116, doi: 10.1007/s11920-019-1094-0.

[21] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Pantic, AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition, *Proceedings of the 9th International on Audio/Visual*

Emotion Challenge and Workshop, Nice France, Oct 15, 2019, 3-12, doi: 10.1145/3347320.3357688.

[22] S. Vashishtha, V. Gupta, M. Mittal, Sentiment analysis using fuzzy logic: A comprehensive literature review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2023, **13**, e1509, doi: 10.1002/widm.1509.

[23] N. Ding, S. Tian, L. Yu, A multimodal fusion method for sarcasm detection based on late fusion, *Multimedia Tools and Applications*, 2022, **81**, 8597-8616, doi: 10.1007/s11042-022-12122-9.

[24] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, W. Gao, Late fusion incomplete multi-view clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **41**, 2410-2423, doi: 10.1109/TPAMI.2018.2879108.

[25] G. D. Jadhav, S. D. Babar, P. N. Mahalle, Exploratory data analysis and algorithmic comparison for early depression detection, *2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, December 13-14, Indore, India, IEEE, 2024, 1-7, doi: 10.1109/ICTBIG64922.2024.10911717.

[26] L. K. Ahire, S. D. Babar, P. N. Mahalle, Fuzzy approach for context identification into ambient computing, *International Journal of Intelligent Systems and Applications in Engineering*, 2023, **11**, 672-681.