Data and Metadata. 2025; 4:863
doi: 10.56294/dm2025863

ORIGINAL

Application of multi-modal data fusion based on deep learning in diagnosis of depression

Aplicación de la fusión multimodal de datos basada en el aprendizaje profundo en el diagnóstico de la depresión

Aimin Pan[1]

[1]College of Computing and Information Technologies, National University. 1008, Manila Philippines.

Cite as: Pan Aimin. Application of multi-modal data fusion based on deep learning in diagnosis of depression. Data and Metadata. 2025; 4:863. https://doi.org/10.56294/dm2025863

Submitted: 08-06-2024 **Revised:** 05-10-2024 **Accepted:** 09-04-2025 **Published:** 10-04-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Aimin Pan

ABSTRACT

Depression is a frequent mental condition requiring precise diagnosis in its early onset. Traditional methods are less than accurate and occur late. Following these deficits, this investigates the multi-modal data fusion and Deep Learning (DL) with the purpose of enhancing accuracy for diagnosis. A new DL model, Dynamic Dolphin Echolocation-tuned Effective Temporal Convolutional Networks (DDE-ETCN), is utilized for depression diagnosis. Different sources of data, such as physiological signals (EEG, heart rate), behavioral indicators (facial expressions), and biometric data (activity levels), are fused. Data preprocessing includes wavelet transformation and normalization of biometric and physiological data, and median filtering of behavioral data to provide smooth inputs. Feature extraction is performed through Fast Fourier Transform (FFT) to obtain frequency-domain features of depression indicators. Feature-level fusion is a good fusion of all data sources, which improves the model's performance. The DDE tuning mechanism improves temporal convolution layers to improve the model's ability in detecting sequential changes. The proposed DDE-ETCN model highly improves depression diagnosis when it is developed in Python. The model attains an RMSE of 3,59 and an MAE of 3,09. It has 98,72 % accuracy, 98,13 % precision, 97,65 % F1-score, and 97,81 % recall, outperforming conventional diagnostic models and other deep learning-based diagnostic models. The outcomes show the efficiency of the model, rendering a more objective and accurate depression diagnosis. Its higher performance justifies its potential for practical use, providing enhanced accuracy and reliability compared to traditional approaches. This innovation emphasizes the necessity of incorporating deep learning for enhanced mental health evaluations.

Keywords: Multi-Modal Data; Deep Learning (DL); Diagnosis; Depression; Dynamic Dolphin Echolocation-Tuned Effective Temporal Convolutional Networks (DDE-ETCN).

RESUMEN

La depresión es una condición mental frecuente que requiere un diagnóstico preciso en su inicio temprano. Los métodos tradicionales son menos precisos y ocurren tarde. Siguiendo estos déficits, se investiga la fusión multimodal de datos y el aprendizaje profundo (DL) con el propósito de mejorar la precisión para el diagnóstico. Un nuevo modelo DL, Dynamic Dolphin Echolocation-tuned Effective Temporal convolunetworks (DDE-ETCN), se utiliza para el diagnóstico de depresión. Se fusionan diferentes fuentes de datos, como señales fisiológicas (EEG, frecuencia cardíaca), indicadores de comportamiento (expresiones faci) y datos biométricos (niveles de actividad). El preprocesamiento de datos incluye la transformación wavelet y la normalización de los datos biométricos y fisiológicos, y el filtrado medio de los datos de comportamiento para proporcionar entradas suaves. La extracción de características se realiza a través de la transformada

rápida de Fourier (FFT) para obtener características de dominio de frecuencia de los indicadores de depresión. La fusión a nivel de características es una buena fusión de todas las fuentes de datos, lo que mejora el rendimiento del modelo. El mecanismo de ajuste DDE mejora las capas de convolución temporal para mejorar la capacidad del modelo en la detección de cambios secuenciales. El modelo propuesto de DDE-ETCN mejora el diagnóstico de depresión cuando se desarrolla en Python. El modelo alcanza un RMSE de 3,59 y un MAE de 3,09. Tiene un 98,72 % de precisión, un 98,13 % de precisión, un 97,65 % de F1-score y un 97,81 % de recuerdo, superando a los modelos de diagnóstico convencionales y otros modelos de diagnóstico basados en el aprendizaje profundo. Los resultados muestran la eficiencia del modelo, haciendo un diagnóstico de depresión más objetivo y preciso. Su mayor rendimiento justifica su potencial para el uso práctico, proporcionando una mayor precisión y fiabilidad en comparación con los enfoques tradicionales. Esta innovación enfatiza la necesidad de incorporar el aprendizaje profundo para mejorar las evaluaciones de salud mental.

Palabras clave: Datos Multimodales; Aprendizaje Profundo (DL); Diagnóstico; Depresión; Redes Convolucionales Temporales Efectivas Sintonizadas con Ecodelación de Delfines Dinámicos (DDE-ETCN).

INTRODUCTION

Depression is established among individuals between the ages of 15 and 25 and is reflected to be a major contributor to disability. Using neuroimaging data, researchers have been identifying and mapping the link between brain shape and function for years.[1] The depressions affect over 300 million people worldwide, as reported by the World Health Organization (WHO). Major psychological suffering, including self-harm and suicidal thoughts, can be caused by depression.[2] Negative schemas about oneself, other people, and situations are connected to depression. It is maintained that depressed individuals in particular are hyper-aware of adverse information and try to find constant negative reactions in societal situations. It's frequently believed that depressed individuals have abnormal facial expression recognition behaviors.[3] Depression is characterized by a low mood and a loss of interest and excitement, but those who are severely depressed might harm themselves or take their lives. To seriously prejudice individuals' lives, depression also troubles society. Reducing the loss of people and society is assisted by early identification of depression. Depression must be recognized early and treated.[4] Interaction between the various types of data, such as text, speech, image, and sensor data, is permitted for multimodal analysis, which uses the information collected to assess the diverse emotions and behaviors of different individuals.[5] There are two types of stress in humans: good and negative. Stress that lasts a brief period when a person's talents are adequate to handle the task is known as positive or severe stress. Stress that persists for a long period when a problem beyond a person's capacity is known as negative or chronic stress. Everyone encounters a difficult situation at some time in their lives and responds appropriately.[6] Heart Rate Variability (HRV), Electro Encephalogram (EEG), and Multimodal Mood Classification (MMC) are three different models and techniques for the determination of an individual's Mood State (MS). It might be tough to combine many multimodal data kinds and show them immediately as an MS. Over the past ten years, high-level multimodal information was efficiently fused.[7] The research objective is to enhance the examination of depression by joining biometric, physiological, EEG, and auditory data in a DL-based multimodal framework.

Fully Connected Heterogeneous Neural Network (FC-HGNN) is used to classify mood states.[8] It functions in two stages: first, a connectome network extracts individual brain properties, followed by multimodal data fusion to produce a heterogeneous population graph. While the approach recognizes key biomarker regions for illness classification, confirmation on larger datasets is necessary for broader generalizability. The combination of Artificial Intelligence (AI) with multimodal physiological inputs is essential for Affective Disorder (AD) detection.[9] Neuromorphic Computing (NC) chips improve wearable diagnostics by providing energy-efficient methods for processing spatial-temporal data. Despite advancements in multimodal AD detection, challenges such as hardware constraints and data heterogeneity continue. The research addresses obtainable progress, challenges, and future directions to support clinical adoption and further research. An enhanced version of the Multimodal Fusion Transformer (MF Former), Spatio-temporal Feature Fusion Transformer (STF2Former), is intended to extract temporal and spatial characteristics from resting-state functional Magnetic Resonance Imaging (rs-fMRI) data for bipolar disorder diagnosis.[10] The STF Aggregation Module (STFAM) enhances multimodal fusion and feature extraction, outperforming MF Former and other advanced methods. Despite its efficiency, further testing on diverse datasets is required, and its computational complexity limits real-time clinical applications. Deep learning (DL) is utilized for depression severity recognition through visual emotion, autonomous auditory, and audiovisual emotion sensing.[11]

A baseline behavioral variable, including speech prosody and facial expressions, was introduced, along with a discussion on trial design and data collection for computer-aided diagnosis. Various baseline attributes were analyzed, contributing to a better understanding of automated depression severity assessments. However, real-world applications and feature generalizability remain challenging, necessitating further clinical validation. For major depressive disorder (MDD) diagnosis, Multi-Stage Graph Fusion Networks (MSGFN) was implemented. [12] Functional connectivity was computed to analyze white and gray matter interactions. A graph convolutional fusion module integrated graphs and features at each stage. The experimental results demonstrated MSGFN's superior performance over existing techniques, though additional verification is needed for clinical applications. A DL-based multimodal system utilizing EEG and audio data was proposed for MDD classification while maintaining privacy.[13] Bi-directional Long Short-Term Memory (Bi-LSTM) was selected as the foundation model due to its superior accuracy for EEG and audio processing. Federated learning demonstrated higher accuracy, preserving data privacy while ensuring reliable MDD classification. However, real-world validation and further optimization are necessary for deployment in edge devices. To overcome troubles with multi-modal feature fusion and long-term context extraction, a design of multi-modal adaptive fusion transformer network is utilized to calculate depression levels.[14]

Audio-visual data was recovered using transformer-based methods, which also incorporated an added depression categorization employment to improve evaluation precision. The method achieved a Concordance Correlation Coefficient (CCC) of 0,733, surpassing the state-of-the-art approach (CCC = 0,696) by 6,2 %. Further refinements are needed for broader practical applications. To address the limitations of the Self-Reported Anxiety Scale (SAS) and Self-Reported Depression Scale (SDS), a multimodal framework was proposed for improved anxiety and depression analysis.[15] The method integrates facial expressions and movement data from video recordings taken while participants complete the scales, enhancing diagnostic accuracy. Combining scale data with facial expressions and movements improves consistency with clinical diagnoses.

The research organization was categorized as following phases, such as phase 2 defined the related research, phase 3 denoted the methodology, phase 4 represented the results and the conclusion was depicted in phase 5.

METHOD

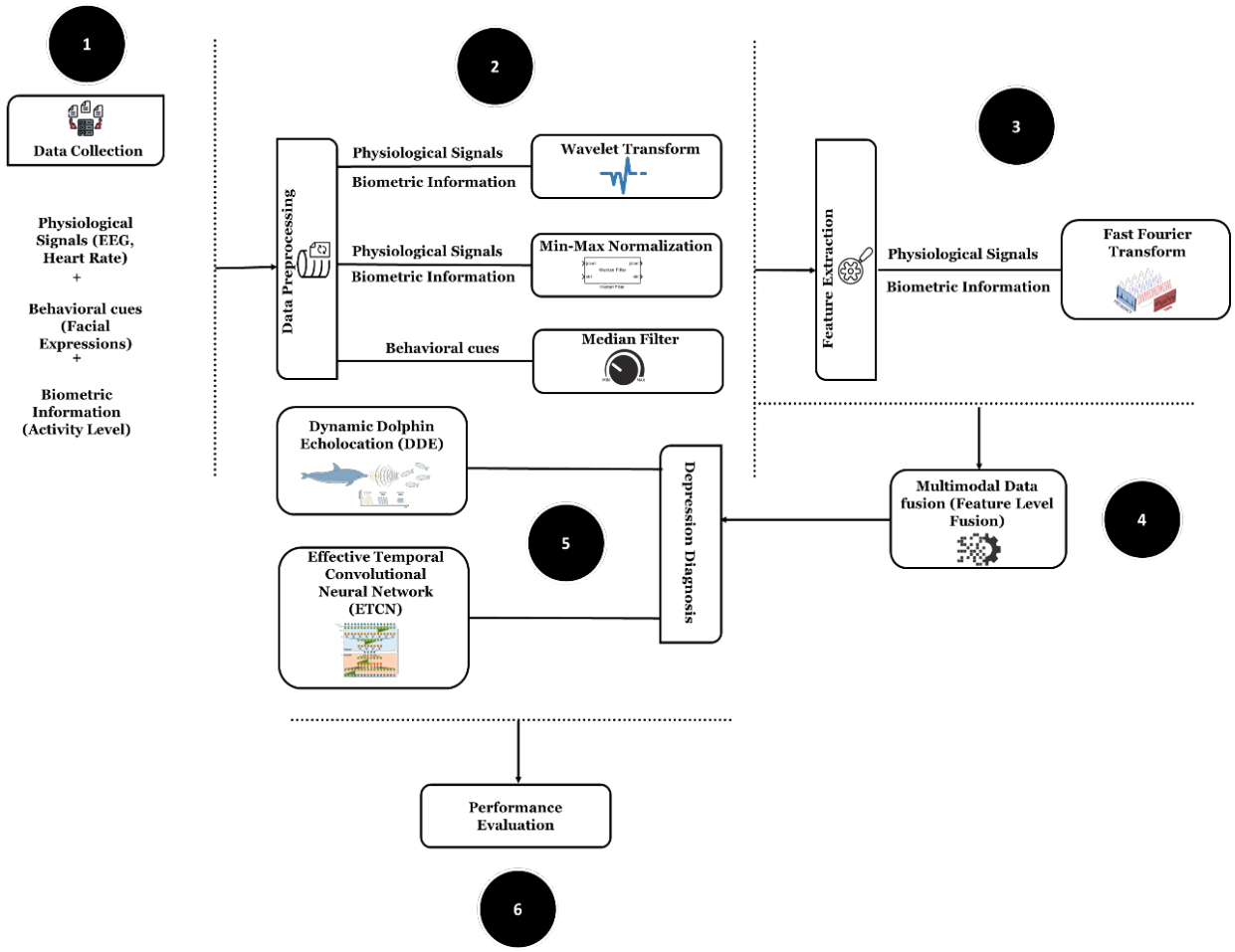

Figure 1. Structure of the proposed methodology

Data gathering from physiological, behavioral, and biometric sources was the measurement of the technique. Wavelet transform, median filtering, and normalization were utilized for preprocessing. While feature-level fusion enhances the analysis of depression, feature extraction applies FFT to physiological, and biometric data and behavioral cues. After that, the effectiveness of the DDE-ETCN model is evaluated and trained. Figure 1 represents the complete process of the enhanced method.

Data Collection

The required data for the diagnosis of depression was collected from the open source called Kaggle: https://www.kaggle.com/datasets/s3programmerlead/multimodal-dataset-for-depression-analysis.

These datasets utilize multimodal data, such as EEG recordings and audio (speech and music), to facilitate emotion recognition and depression diagnosis. To examine expressive and cognitive patterns to provide vital resources for AI-driven mental health evaluation was completed by the dataset. The programs enhance models for repeated and more precise diagnosis.

Data Preprocessing

The development of revolving raw data into a clear, organized, and useful format for examination is known as data preparation. To enhance signal clarity, wavelet transform and normalization were used in data preparation for physiological and biometric data for depression diagnosis. To remove noise and outliers from behavioral data, especially facial expressions, median filtering is used. By standardizing feature ranges, data normalization guarantees consistency.

Wavelet Transform (WT)

A popular data preparation technique for attaining signal characteristics in the frequency and temporal domains is WT. It generates time-frequency images that represent signal properties by converting the raw signals into time-frequency distributions. An interior operation using WT was executed on the signal and a wavelet family was obtained by scaling and translating a mother wavelet. The feature set was advanced and improved by the integration of physiological and biometric data, like Heart Rate Variability (HRV), activity levels. WT is used in this research for DL-based depression diagnosis through multimodal data fusion. The approach enhances feature extraction by converting auditory and EEG inputs into time-frequency representations, which elevates the robustness and accuracy of analysis. The mother wavelet is represented by equation (1).

$$WT = \psi_{t,s} \frac{1}{\sqrt{t}} \left(\frac{s-S}{t} \right) \qquad (1)$$

Where represents the parameter of translation, is the scale parameter, has an inverse relationship with frequency. The following mathematical classification (equation 2), can be performed to establish the WT of a given signal y(s).

$$W(t,s) = \langle y(s), \psi_{t,s} \rangle = \frac{1}{\sqrt{t}} \int y(s)\, \psi * \left(\frac{s-S}{t} \right) d(t) \qquad (2)$$

Where the complex conjugate was represented by ψ*.This process evaluates the WT to the Fourier transform, which crack down a signal into its entity frequencies. Using the wavelet family, this equation breaks down the signal y(s) into a collection of wavelet coefficients. According to the abovementioned equations, family wavelets have two different parameters: t ands. By utilizing the potential, the research means to improve the diagnosis of depression by integrating biometric, EEG, aural data, and psychological data in a DL-based multimodal framework. The signal y(s)is transformed by the family wavelets following the convolution process, and it is expected to the two-dimensional (2-D) time and frequency dimensions. Temporal-frequency images are generated from one-dimensional temporal data. The selected scales are associated to the frequency range.

Min-Max Normalization

A preprocessing method called normalization regulates feature values to maintain relationships and guarantee consistency across various data formats. Normalization develops the integration of multimodal data for analyzing depression, which amplifies the model's capacity to successfully categorize patterns linked to depression. In the proposed approach for depression diagnosis, min-max normalization is engaged to preprocess physiological and biometric data and make certain consistency across multiple modalities. This technique scales feature values with a predefined range, typically [0, 1], preserving the intrinsic interaction within the data. The transformation follows the equation (3).

https://doi.org/10.56294/dm2025863

$$u' = \frac{u - min_B}{max_B - min_B}(new\_max_B - new\_min_B) + new\_min_B \quad (3)$$

The normalized value was signified by u', the imaginative feature value was symbolized by u', the maximum and minimum values of feature B were designated by max_B and min_B, and new_max_B together with new_min_B provided the normalization goal range. By optimizing the combination of multimodal data, this preprocessing step improves the model's capacity to identify patterns associated with depression.

Median Filtering

A non-linear filtering technique called the median filter generates the output by scheming the median of the input values within a secured window size. A spatial median filter was used for two-dimensional data, such as image structures of facial expressions, whereas an equivalent window was engaged for one-dimensional behavioral data, such as face expression sequences. The median filter was utilized in the research to eliminate noise and outliers from behavioral data, resulting in more reliable and responsible facial expression inputs for the analysis of depression. By removing salt-and-pepper noise and supporting important structural information, the filtering process improves the ability of features utilized in the recommended model. The median filter was computed using the equation (4).

$$Img(w,z) = median\{Img_{act}((w+j, z+i) | j,i \in Q)\} \quad (4)$$

Where Img(w,z) characterizes the output image of the filter at position (w,z), Img_{act} (w+j, z+i) was the real image values, j,i∈Q represents the adjoining pixels and the median computes the median values of all pixels.

Feature Extraction using Fast Fourier Transform

Feature extraction extracts important patterns from multi-modal data sources, was necessary to enhance the diagnosis accuracy of depression. To transform time-domain signals into frequency-domain representations, FFT is utilized in this research on physiological and biometric signals, such as heart rate, activity levels, and EEG. This adjustment makes it possible to recognize the prominent frequency components associated with physiological changes conveyed by depression. The process involved in the FFT is given below:

1. Fourier coefficient calculation: the Fourier coefficients of a specified signal x(t) in the frequency range [0,2π] were at tained through the FFT algorithm. The Fourier transform was represented in equation (5).

$$E_q = \sum_{v=0}^{n-1} w_m f^{-j2\pi\frac{v}{n}} \quad (5)$$

Where E_q depicts the frequency domain coefficients, and n is the length of the input signal.

2. Magnitude Computation: the entire values of the computed coefficients are achieved as $B_q = |E_q|$, ensuring robust feature representation.

3. Matrix Representation: the removed frequency-domain features are permitted into an ordered n×m matrix, assist well-organized input into DL models. The transformation is followed in equation (6).

$$W = \begin{bmatrix} B_1 B_2 \ldots\ldots B_m \\ B_{m+1} B_{m+2} \ldots\ldots B_{m+m} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ B_{(n-1)(m+1)} B_{(n-1)(m+2)} \ldots\ldots B_{(n-1)(m+m)} \end{bmatrix}, q = n \times m \quad (6)$$

Where m and n signify the columns and rows of the matrix, respectively. The applied FFT-based feature extraction allows the DDE-ETCN model to efficiently identify depression-related signal variations, enhancing diagnostic accuracy by leveraging frequency-domain characteristics.

Multimodal Data Fusion based depression diagnosis

To increase the accuracy and reliability of diagnosing depression, a DL architecture that combines multimodal data fusion was essential. Conventional diagnostic methods repeatedly rely on subjective, one-modal data, including self-reported feedback or clinical assessments, which are exposed to prejudice and unpredictability. It is possible to get a more methodical and impartial valuation of depressed trends by methodically integrating behavioral, biometric, and psychological data.

https://doi.org/10.56294/dm2025863

Psychological Data: this includes self-reported cognitive evaluations, mood states, and emotional reactions. These components supply an improved acceptance of depression tendencies by offering an approach to mental health state and cognitive processes.

Biometric Data: neurophysiological alterations associated to depression were captured by physiological signals, such as EEG measurements, HRV, and galvanic skin response (GSR). These intentional assessments are significant markers of emotional deregulation and stress.

Behavioral Information: speech, facial expressions, sleep patterns, and social interactions can show signs of depression. This technique was useful for premature detection because subtle behavioral changes frequently occur before clinical diagnosis.

The DL models develop feature extraction and pattern recognition by integrating all three data sources, which elevates the classification accuracy and resilience of depression. The combined method supports more accurate and prompt depression diagnosis by reducing biases, ensuring data integrity, and improving predictive capacities.

Dynamic Dolphin Echolocation-tuned Effective Temporal Convolutional Networks (DDE-ETCN) for the diagnosis of depression

The DDE-ETCN model is recommended for specific depression identification. This technique integrates temporal CNN with optimization stimulated by dolphin echolocation to get better feature extraction and classification. The evaluation enhances detection accuracy by efficiently capturing slight behavioral indicators and emotional fluctuations. The recommended approach dynamically optimizes network parameters to conquer the limitations of current models. The improved framework affords a dependable and flexible way to identify depression early and accurately.

Effective Temporal Convolutional Neural Network

A proficient technique for time-series prediction, (TCNs) was particularly useful for applications like voice modeling and human activity detection. TCNs' primary elements were residual connections, causal convolutions, and dilated convolutions, which are exceptional for use, such as the real-time recognition of depression. To enhance long-term dependency representation while preserving computing efficiency, the ETCN was created. The technique was especially helpful for tasks like speech-to-speech depression detection, where precise predictions depend on sequential dependencies and minute changes in voice features. To enhance learning stability and predictive performance, the ETCN combines residual connections, dilated convolutions, and causal convolutions.

Causal convolution

To avoid future information from affecting the present time step, causal convolution that forecast at time t is exclusively based on historical data. For real-time depression identification, where recent and previous speech data only available, and this is essential for the detection. Here's the causal convolution operation depicted in equation (7).

$$E(w_s) = \sum_{j=1}^{l} E_i \, w_s - l + j \qquad (7)$$

Where the voice signal input at time E is denoted by w_s and E stands for the convolution kernel of size l. The sequential integrity of speech components is sustained by this configuration, which is necessary for examining temporal patterns in the identification of depression.

Dilated convolution

A minimal increase in network depth and dilated convolutions allows the model to capture long-term dependencies in speech signals, which is especially functional for detecting depression because emotional and prosodic variations occur over long periods. The receptive field Z is given by the equation (8).

$$Z = \sum_{m=1}^{M} (l-1).c_m + 1 \qquad (8)$$

Where M is the number of layers, l is the size of the kernel, and c is the exponentially growing dilation factor. The ETCN model learns pertinent speech characteristics effectively and with minimal computing cost by employing dilation.

Residual connections

ETCN utilizes residual connections to improve training efficiency and stability by dropping problems, like disappearing and bursting gradients. Deeper structures are prepared and made feasible by these connections

https://doi.org/10.56294/dm2025863

without sacrificing efficiency, which makes them useful for intricate sequential tasks like depression detection. The representation of the residual block is shown in equation (9).

$$P = Activation\big(E(w) + conv1\ C(w)\big) \qquad (9)$$

Where E(w) and conv1 C(w)designate the transformation operations applied to the input speech features and the 1D convolution in the residual path, respectively. ETCN uses residual learning to guarantee reliable feature extraction from speech data, increasing the precision of depression prediction models.

Dynamic Dolphin Echolocation

Dolphins utilize echolocation, which absorbs making ultrasonic clicks and examines the echoes that return, to find the way and locate objects in the environment. To exploit sequential data processing for depression detection, DDE approach minimizes the detection range according to received signals. The DDE technique suggests how dolphins gradually sharpen their search. The Convergence Factor (CF), which energetically modifies the search space, was a mathematical representation of the procedure represented in equation (10).

$$CF_i = \frac{UL_i - LL_i}{2SD_i} \qquad (10)$$

Where CF_i represents the convergence factor for the variablei, UL_i-LL_i depicts the upper and lower limits of variable i, and $2SD_i$ represents the standard deviation of location values. The approach can dynamically regulate its feature extraction progression, which also amplifies the sensitivity to tiny changes in speech and physiological data, thusenhancing the classification of depression. By iteratively fine-tuning weights, the model optimally discovers temporal relationships, much like dolphins adjust their echolocation signals to accurately identify targets. DDE optimization was integrated into the recommended framework to efficiently classify behavioral and emotional indicators connected to depression. Improved accuracy and robustness in AI-driven mental health evaluation are assured by this method. The DDE-ETCN method improves feature extraction and pattern identification by the combination of effective temporal convolutional networks with dynamic dolphin echolocation-based optimization. By optimizing multimodal data processing, this fusion enhances the precision and dependability of depression diagnosis.

RESULTS

The DDE-ETCN model performed exceptionally well in accuracy, precision, recall, and F1-score when tested for depression diagnosis with error estimation using MSE and RMSE. Feature extraction and sequential pattern recognition were enhanced by combining multi-modal data fusion, and DDE-based optimization, which improved diagnostic dependability.

Experimental setup

Python 3.8 was utilized to develop the implemented technique with an Intel i7 CPU running on Windows 11. Some libraries were used for DL. TensorFlow, and PyTorch were utilized for model implementation and behavioral data analysis, an OpenCV was used. Matplotlib and Seaborn were used to interpret the model's evaluation metrics.

Performance Analysis

The DDE-ETCN technique's performance displays the function of multi-modal data fusion to identify depression. The technique improves its capability to acquire a range of depression-related characteristics by integrating physiological signals (heart rate, EEG), biometric data (activity levels), and behavioral cues (facial expressions). The technique outperforms conventional diagnostic methods in terms of prediction accuracy and dependability by integrating many modalities. The results emphasize how crucial it is to combine different kinds of data to enhance the model performance.

Table 1. Comparison of performance metrics across modalities

Modality	Accuracy (%)	Precision (%)	F1-Score (%)	Recall (%)
Psychological signal	98,55	97,45	97	97,12
Behavioral Cues	98	97	96,19	96,55
Biometric Information	97	96,55	96	96,12
Multi-Modal fusion (DDE-ETCN)	98,72	98,13	97,65	97,81

https://doi.org/10.56294/dm2025863

The DDE-ETCN technique metrics were evaluated across several modalities in table 1. These measurements show that the model works with multi-modal data to identify depression.

Accuracy: the accuracy of the technique, which determines the percentage of correct predictions, determines its overall performance. The accuracy of the depression, diagnostic technique measures how effectively the system recognizes cases of depression guaranteeing accurate and successful predictions based on multi-modal data. The depiction of accuracy in figure (2) over different factors was illustrated.

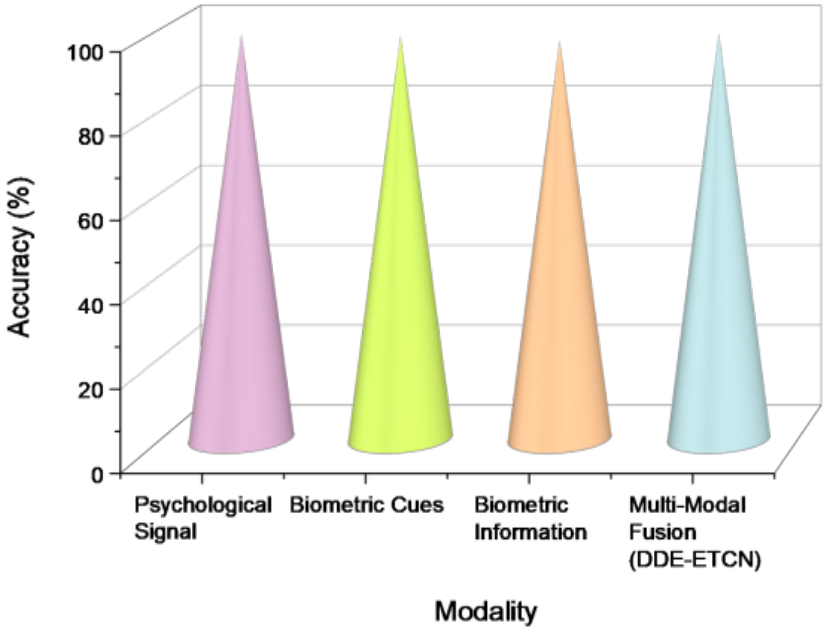

Figure 2. The graphical representation of accuracy

Precision: precision evaluates the percentage of actual positive diagnoses for depression out of all the estimated positives. Precision lowers the possibility of false positives and increases dependability in the model by undertaking that a diagnosis was made with a high probability of precision. Figure 3 shows the illustration of precision over various parameters.

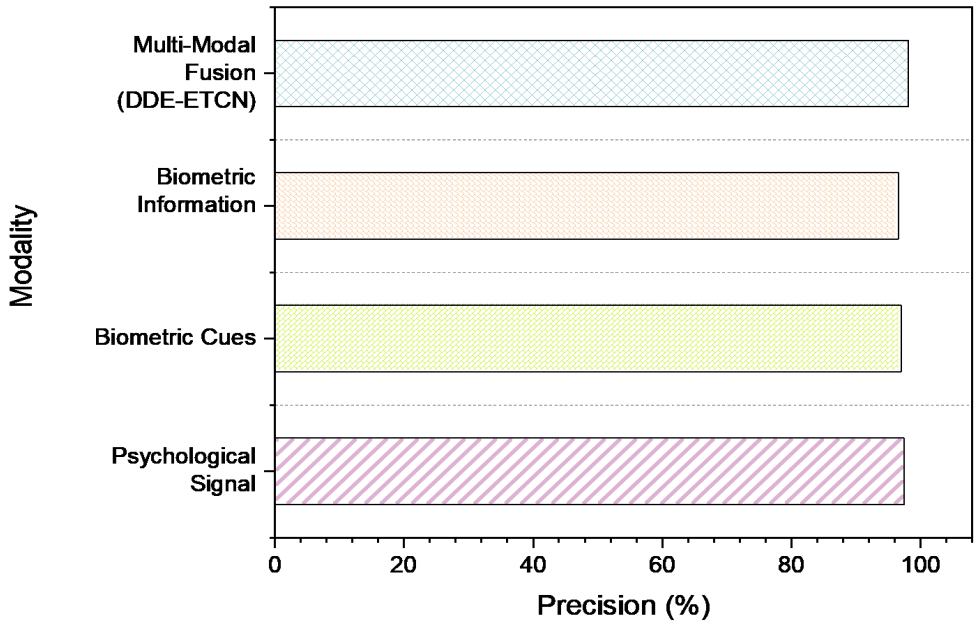

Figure 3. The graphical representation of Precision

Recall: a recall in this technique ensures that the structure constantly identifies depression, reducing false negatives and makes certain rapid interference. The model's recall estimates how well it can identify depressive symptoms. Figure 4 shows the compared recall values.

https://doi.org/10.56294/dm2025863

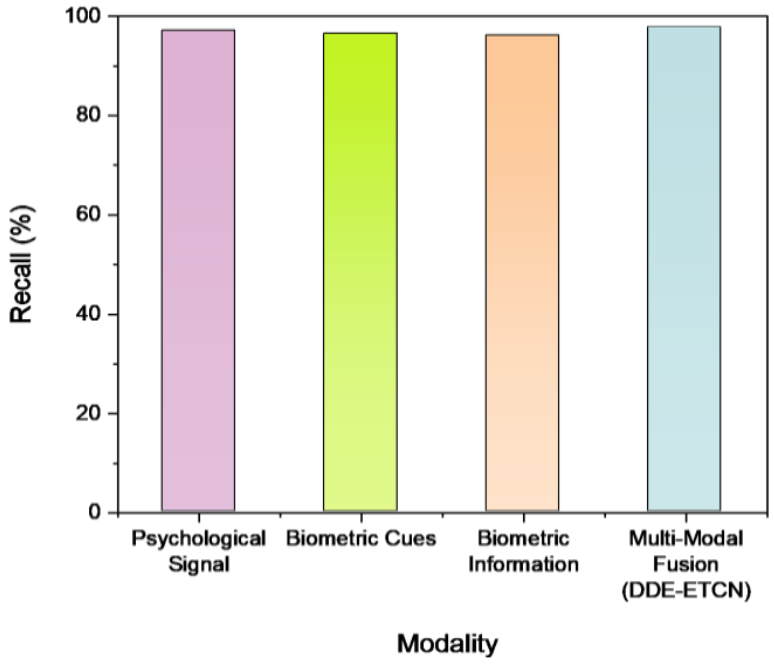

Figure 4. The graphical representation of recall

F1-Score: the accuracy and recall were stabilized with the help of F1-Score that was used to analyze the model's overall performance. It assured both accuracy and sensitivity when detecting reserved depression signals from multi-modal data, making it more supportive in assessing the DDE-ETCN strategy for depression diagnosis. Figure 5 depicts the F1-Score comparison over various modalities.

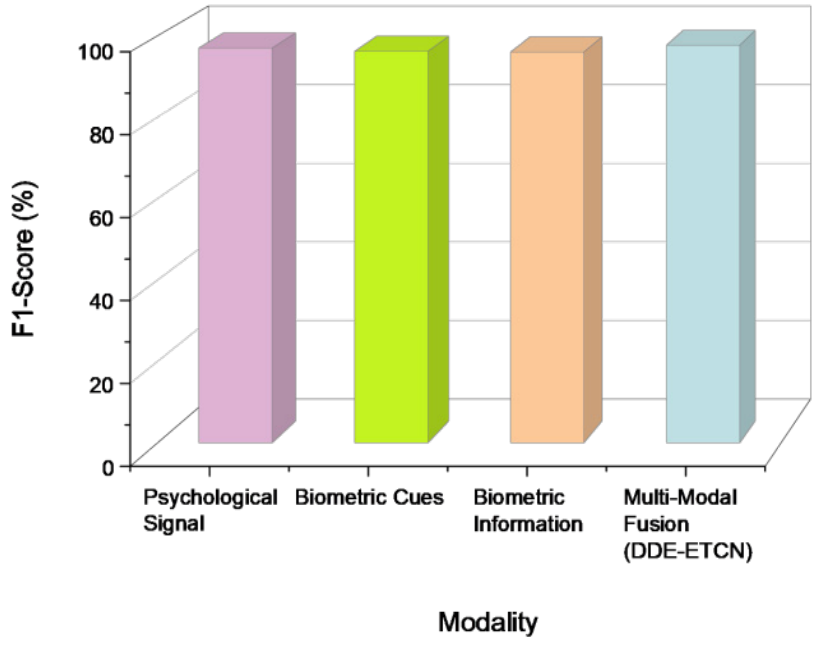

Figure 5. Graphical representation of F1-Score

Evaluation of error metrics

Multi-Feature Modulation with Attention (MFM-Att) enhances feature extraction by dynamically selecting relevant features, but its dependence on single-modal data limits its effectiveness in capturing depression-related patterns.[16] The implemented method overcomes by incorporating multi modal data, ensuring a more inclusive demonstration of depressive states. MAE and RMSE were used in the evaluation to assess the DDE-ETCN model's performance. With the use of multi-modal data, such as biometric data, physiological signals, and behavioral cues, these metrics assist in calculating the model's prediction accuracy for diagnosing depression. By leveraging various data sources, DDE-ETCN reduces prediction errors more effectively than the existing approach, enhancing model reliability in real-world applications. Through a comparison of the error metrics,

https://doi.org/10.56294/dm2025863

the measurement shows the model that decreases prediction errors and guarantees reliability in practical applications. Table 2 depicts the comparative analysis of error metrics across methods.

Table 2. Error Metrics comparison for the model's performance

Methods	MAE	RMSE
MFM-Att[16]	3,18	3,68
DDE-ETCN [Proposed]	3,09	3,59

MAE: a standard absolute feature between the designed values and actual values was developed to examine the MAE measure, which approximates the depression detection technique's accuracy. When analyze a depression, a lesser MAE shows that the model's predictions were concatenated in line with the actual values, signifying the implication of multi-modal data fusion and the recommended DL method for precise depression identification. Figure 6 depicts the graphical representation of MAE.

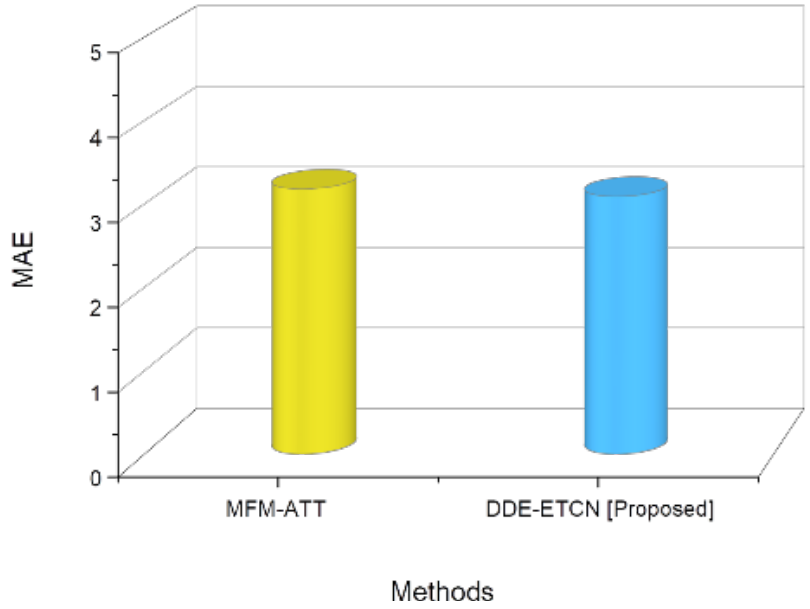

Figure 6. Graphical representation of MAE over methods

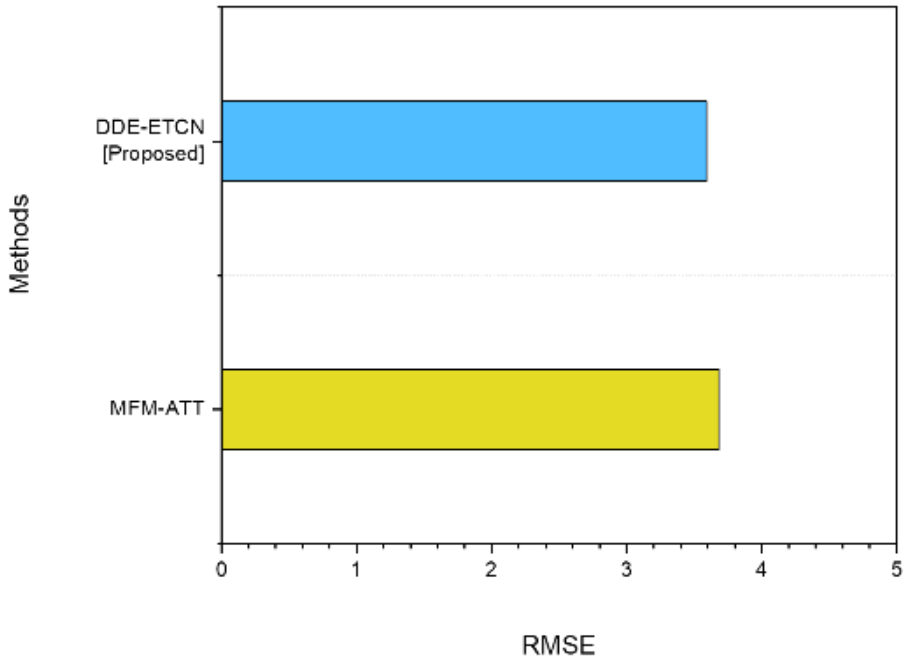

Figure 7. Graphical Representation of RSME

RMSE: an important measure for evaluating depression detection algorithms efforts was RMSE, which illustrates the difference between predicted and actual results. Better prediction accuracy in the classification of depression based on multi-modal data, such as behavioral cues, physiological signals, and biometric information, is shown in figure 7. The DDE-ETCN representations were utilized to reduce RMSE to enhance clinical decision-making by increasing the reliability of depression diagnosis.

DISCUSSION

To enhance the accuracy and reliability of diagnosing depression by the integration of physiological signals (heart rate, EEG), behavioral cues (facial expressions), and biometric data (activity levels). The analytical performance of existing approaches was generally insufficient due to the limitations in managing multi-modal data integration. By integrating temporal convolution with dynamic tuning, the recommended DDE-ETCN technique effectively gets beyond these limitations and guarantees precise depression detection. The DDE-ETCN model outperformed existing multimodal data fusion approaches, achieving higher accuracy (98,72 %), precision (98,13 %), F1-score (97,65 %), and recall (97,81 %), according to tentative data, which showed distinguished gains. The technique established assurance for practical clinical applications by outperforming conventional approaches in error measures similar to MAE (3,09) and RMSE (3,59).

Limitation and Future Scope

The executed DDE-ETCN techniques also have some drawbacks but demonstrate major developments in the diagnosis of depression. It demands a vast amount of computer power because of its difficulty, and its achievement is reliant on the quality and accessibility of multi-modal data, which is not available at all times. Additionally, the model might be influenced by data irregularity and noise. Using noise-reduction approaches to improve the model's resilience, making it more capable for real-time application, and expanding its applicability to more varied datasets that might be the major purpose of future research. Additionally, investigating more problematical DL techniques, such as reinforcement learning, might increase the model's adaptability.

BIBLIOGRAPHIC REFERENCES

1. Meshram, P. and Rambola, R.K., 2023. RETRACTED: Diagnosis of depression level using multimodal approaches using deep learning techniques with multiple selective features. Expert Systems, 40(4), p.e12933. https://doi.org/10.1111/exsy.12933

2. Wang, Y., Wang, Z., Li, C., Zhang, Y. and Wang, H., 2022. Online social network individual depression detection using a multitask heterogenous modality fusion approach. Information Sciences, 609, pp.727-749. https://doi.org/10.1016/j.ins.2022.07.109

3. Rajawat, A.S., Bedi, P., Goyal, S.B., Bhaladhare, P., Aggarwal, A. and Singhal, R.S., 2023. Fusion fuzzy logic and deep learning for depression detection using facial expressions. Procedia Computer Science, 218, pp.2795-2805. https://doi.org/10.1016/j.procs.2023.01.251

4. Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., Li, Y. and Hu, B., 2022. TAMFN: time-aware attention multimodal fusion network for depression detection. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, pp.669-679. https://doi.org/10.1109/TNSRE.2022.3224135

5. Patil, M., Mukherji, P. and Wadhai, V., 2024. Federated learning and deep learning framework for mri image and speech signal-based multi-modal depression detection. Computational Biology and Chemistry, 113, p.108232. https://doi.org/10.1016/j.compbiolchem.2024.108232

6. Kuttala, R., Subramanian, R. and Oruganti, V.R.M., 2023. Multimodal hierarchical CNN feature fusion for stress detection. IEEE Access, 11, pp.6867-6878. https://doi.org/10.1109/ACCESS.2023.3237545

7. Tai, C.H., Chung, K.H., Teng, Y.W., Shu, F.M. and Chang, Y.S., 2021. Inference of mood state indices by using a multimodal high-level information fusion technique. IEEE Access, 9, pp.61256-61268. https://doi.org/10.1109/ACCESS.2021.3073733

8. Gu, Y., Peng, S., Li, Y., Gao, L. and Dong, Y., 2025. FC-HGNN: A heterogeneous graph neural network based on brain functional connectivity for mental disorder identification. Information Fusion, 113, p.102619. https://doi.org/10.1016/j.inffus.2024.102619

9. Tian, F., Zhang, L., Zhu, L., Zhao, M., Liu, J., Dong, Q. and Zhao, Q., 2024. Advancements in affective

disorder detection: Using multimodal physiological signals and neuromorphic computing based on snns. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2024.3420445

10. Wang, G., Fan, F., Shi, S., An, S., Cao, X., Ge, W., Yu, F., Wang, Q., Han, X., Tan, S. and Tan, Y., 2024. Multi-modality fusion transformer with spatio-temporal feature aggregation module for psychiatric disorder diagnosis. Computerized Medical Imaging and Graphics, 114, p.102368. https://doi.org/10.1016/j.compmedimag.2024.102368

11. Liu, J., Huang, Y., Chai, S., Sun, H., Huang, X., Lin, L. and Chen, Y.W., 2022. Computer-aided detection of depressive severity using multimodal behavioral data. Handbook of Artificial Intelligence in Healthcare: Vol. 1-Advances and Applications, pp.353-371. https://doi.org/10.1007/978-3-030-79161-2_14

12. Kong, Y., Niu, S., Gao, H., Yue, Y., Shu, H., Xie, C., Zhang, Z. and Yuan, Y., 2022. Multi-stage graph fusion networks for major depressive disorder diagnosis. IEEE Transactions on Affective Computing, 13(4), pp.1917-1928. https://doi.org/10.1109/TAFFC.2022.3205652

13. Gupta, C., Khullar, V., Goyal, N., Saini, K., Baniwal, R., Kumar, S. and Rastogi, R., 2023. Cross-Silo, Privacy-Preserving, and Lightweight Federated Multimodal System for the Identification of Major Depressive Disorder Using Audio and Electroencephalogram. Diagnostics, 14(1), p.43. https://doi.org/10.3390/diagnostics14010043

14. Sun, H., Liu, J., Chai, S., Qiu, Z., Lin, L., Huang, X. and Chen, Y., 2021. Multi-modal adaptive fusion transformer network for the estimation of depression level. Sensors, 21(14), p.4764. https://doi.org/10.3390/s21144764

15. Xie, W., Wang, C., Lin, Z., Luo, X., Chen, W., Xu, M., Liang, L., Liu, X., Wang, Y., Luo, H. and Cheng, M., 2022. Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model. Computerized Medical Imaging and Graphics, 102, p.102128. https://doi.org/10.1016/j.compmedimag.2022.102128

16. Fang, M., Peng, S., Liang, Y., Hung, C.C. and Liu, S., 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control, 82, p.104561. https://doi.org/10.1016/j.bspc.2022.104561

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Data curation: Aimin Pan.
Methodology: Aimin Pan.
Software: Aimin Pan.
Drafting - original draft: Aimin Pan.
Writing - proofreading and editing: Aimin Pan.