Article Type: Research Article

Journal of Neurology and Neurological Sciences

# Patient-centric multi-modal major depressive disorder detection

## Abstract

Major Depressive Disorder (MDD) remains a pressing global health issue, necessitating diagnostic models that effectively capture both neural and emotional cues. This study proposes a multi-modal framework that integrates generalised Partial Directed Coherence (gPDC) measures from EEG signals and LOG-BASED MEL SPECTROGRAM (LBMS) images from speech data. Two modality-specific encoders, each incorporating Convolutional Block Attention Mechanisms (CBAMs), extract neurophysiological and acoustic features, which are fused using a cross-modal attention mechanism to capture inter-modality dependencies.

A rigorous patient-centric data splitting strategy is employed to mitigate data leakage and ensure reliable generalisation. EEG and audio data from the same participant are kept within the same fold, preserving their natural correlation. The proposed model achieves 97.86% accuracy—outperforming previous patient-centric approaches—demonstrating its effectiveness and clinical potential in MDD detection.

**Sandura Shumba*; Johannes Coetzer**
*Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa.*

***Corresponding author: Sandura Shumba**
Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa.
Email: shumba_s@yahoo.com

**Received:** Oct 25, 2025; **Accepted:** Dec 22, 2025; **Published:** Dec 29, 2025

Journal of Neurology and Neurological Sciences
Volume 1 Issue 2 - 2025
www.jnans.org


**Citation:** Shumba S, Coetzer J. Patient-centric multi-modal major depressive disorder detection. J Neurol Neuro Sci. 2025; 1(2): 1014.

**Keywords:** MDD detection; Cross-modal attention; Multi-modal.

## Background and motivation

Major Depressive Disorder (MDD) is a pervasive and debilitating mental illness, currently affecting over 350 million individuals worldwide. According to forecasts by the World Health Organization, MDD is projected to become the leading cause of disease burden by 2030 (World Federation for Mental Health 2012; World Health Organization and others 2017). Particularly concerning is its rising prevalence among children and adolescents, now estimated to affect nearly 30 million young people globally [15]. Despite the scale of the issue, it is estimated that approximately 80% of mental health disorders—including depression—remain undiagnosed (Jordan 2012), highlighting an urgent need for early detection and more objective diagnostic tools.

At present, depression is primarily diagnosed through self-report questionnaires and structured clinical interviews [5,30]. While widely adopted in clinical practice, these approaches are inherently subjective, relying heavily on patients' ability to accurately report their internal states. Many individuals may struggle to articulate their symptoms or may underreport them, leading to under diagnosis or misdiagnosis. Given the substantial impact of MDD on quality of life and everyday functioning, it is essential to develop more reliable, data-driven diagnostic methods (Jordan 2012).

Electroencephalography (EEG) has gained traction as a non-invasive and cost-effective tool for assessing neural activity. A growing body of evidence supports its utility in identifying neuropsychiatric conditions, including MDD [24] (Saeidi et al. 2021; Sharma et al. 2018). EEG enables the real-time capture of brain dynamics, offering insights into both normal and pathological neural processes [16]. Studies have reported distinctive EEG patterns and regional alterations in individuals with depression [7,25,26,28].

In parallel, speech has emerged as another promising biomarker for MDD. Individuals with depression often exhibit characteristic vocal traits, such as reduced pitch variability, slower speech rate, lower intensity, and increased pause duration [8,27] (Mundt et al. 2012). These features have

prompted increasing interest in speech-based automatic depression detection systems [11,17,22,28,29].

The advent of deep learning has further advanced the field of biomedical signal analysis. These models are capable of learning complex, hierarchical representations from raw input data, offering notable improvements over traditional methods (Zhang et al. 2015). While significant progress has been made using deep learning on EEG or speech data independently, a multimodal approach may provide a more comprehensive understanding of depressive states. EEG captures neural correlations of mental processes, whereas speech reflects outward emotional and behavioural expression. Combined, these modalities offer complementary perspectives that may enhance the accuracy and reliability of automated MDD detection systems.

**Related work**

While recent advances in machine learning have shown promise in automating depression detection, most existing approaches still face critical limitations. This section reviews four key research areas relevant to this study: (1) unimodal EEG- and audio-based methods, (2) attention mechanisms, (3) cross-modal attention, and (4) multimodal fusion frameworks. By identifying gaps across these domains—particularly in modality integration, attention design, and evaluation protocols—this review motivates the need for a more comprehensive and interpretable multimodal approach.

**EEG-based and audio-based approaches**

The introduction of deep learning allowed models to learn representations directly from raw or transformed EEG inputs. CNN and hybrid CNN-LSTM models, such as those proposed by Acharya et al. [1] and Ay et al. [4], reported accuracies exceeding 99%. Other work explored alternative inputs, including STFT images and connectivity matrices (Rafiei and Wang 2022; Saeedi et al. 2021), with encouraging results. Nonetheless, many of these studies suffer from methodological issues such as small sample sizes, subject overlap between training and test sets, and insufficient regularisation—all of which compromise their clinical applicability (Xia et al. 2023).

In parallel, audio-based approaches have gained traction due to their non-invasive, accessible nature [11,17]. Early efforts used shallow classifiers like SVMs and GMMs with prosodic or MFCC features (Long et al. 2017) [23], followed by decision-tree-based models [28]. More recent work has embraced deep learning, with CNNs and RNNs applied to log-spectrograms and temporal acoustic features. For instance, Dubagunta et al. [14] leveraged CNNs to detect vocal fold cues of depression, while Wang et al. [19] proposed an attention-enhanced 3D-CBHGA model, achieving 77.14% accuracy. However, these models generally underperform compared to EEG-based systems.

Despite lower standalone accuracy, audio-based methods offer practical advantages and capture behavioural cues that EEG may miss. Their unobtrusiveness and ease of acquisition make them ideal candidates for integration in multimodal systems, where they can complement the neurophysiological insights provided by EEG.

Taken together, these trends underscore the need for multimodal deep learning approaches that combine EEG and audio modalities. By fusing physiological and behavioural data, such systems can address the limitations of unimodal models and support more generalisable, and clinically applicable tools for MDD detection.

**Attention mechanisms in depression detection**

Attention mechanisms have transformed deep learning by enabling models to prioritise salient features in complex data. In MDD detection, however, their use remains limited and often generic. Many studies incorporate attention layers without adapting them to the structural or contextual characteristics of the modality.

For instance, hierarchical attention models have been applied to textual data [29] (Xezonaki et al. 2020), while simple attention mechanisms have been used in EEG and facial data analyses [19,28]. A relevant example is the model proposed by Wang et al. (2022), which combines 1D-CNNs with GRUs and attention layers to dynamically weight EEG features, yielding strong performance.

Nevertheless, these implementations typically treat attention as an auxiliary module rather than an integrated, modality-sensitive component. Few models attempt to harness EEG's temporal, spectral, or spatial characteristics in a learned and explainable manner. This stands in contrast to domains such as Natural Language Processing (NLP) and computer vision, where attention mechanisms are increasingly tailored to modality-specific structures.

**Cross-modal attention**

Cross-modal attention mechanisms offer a promising yet under-explored opportunity for MDD detection. Unlike traditional fusion techniques—such as feature concatenation or decision-level integration—cross-modal attention enables one modality to condition or guide the representation of another. This is especially relevant when combining EEG and audio data, which provide temporally aligned but qualitatively distinct information streams.

In other domains, cross-modal attention has been successfully applied to tasks including video-audio alignment (Min et al. 2021), multimodal anomaly detection [13], and action recognition (Tsai and Chu 2022). These studies demonstrate not only improved performance but also enhance interpretability and data efficiency.

Despite these advantages, cross-modal attention remains virtually absent in MDD detection. Existing models tend to apply attention independently within each modality or rely on naïve fusion strategies, thereby neglecting the nuanced, bidirectional relationships between EEG and speech signals. This omission is significant, as depressive symptoms often manifest simultaneously across multiple modalities—such as altered brain activity occurring alongside atypical speech patterns.

**Multimodal fusion approaches**

Multimodal systems integrating EEG, speech, facial, and textual data have shown enhanced accuracy compared to unimodal counterparts. Studies such as Williamson et al. (2016), Zhao et al. (2021), and Jung and Kim (2020) illustrate the benefits of combining diverse physiological and behavioural signals. However, several critical limitations persist.

Firstly, most fusion strategies do not explicitly model inter-modality dynamics. Attention mechanisms, where used, are typically restricted to single modalities. Secondly, data augmentation tailored to specific modalities is often lacking, making

models prone to overfitting. Thirdly, many studies incorrectly split recordings from the same subjects across training and test sets, undermining clinical validity (Xia et al. 2023).

One noteworthy exception is the MS2-GNN model by Chen et al. (2022), which separates modal-shared and modal-specific embeddings and enforces patient-level data separation. While it marks an improvement, the absence of cross-modal attention limits its ability to capture deeper interactions between modalities.

Recent studies—such as those by Gupta et al. [18], Qayyum et al. (2023), and Zheng et al. (2023)—have employed Vision Transformers, Bi-LSTM pipelines, and knowledge-guided frameworks for multimodal MDD detection. Yet, despite their architectural sophistication, these models also fail to exploit cross-modal relationships effectively, with attention mechanisms confined to intra-modal operations.

Similarly, work by Hu et al. [21] using Large Language Models (LLMs) for EEG and audio interpretation achieved only modest improvements, constrained by the limitations of zero-shot prompting and a lack of modality-specific tuning.

This study directly addresses three major limitations in the current literature. First, conventional fusion strategies are replaced with an adaptive cross-modal attention framework that learns diagnostic dependencies between EEG spectral dynamics and vocal patterns. Second, attention mechanisms are implemented to process spatial-channel interactions in EEG and temporal features in speech. Third, rigorous subject-level data separation and modality-specific augmentation are enforced to ensure clinically interpretable and statistically sound results.

## Contributions

This study offers two key contributions to the field of depression detection using multi-modal data:

1. **A multi-modal diagnostic model with cross-modal attention:** A deep learning architecture is proposed that fuses generalised Partial Directed Coherence (gPDC)-based EEG images and Log-Based Mel Spectrograms (LBMS) from speech, via a cross-modal attention mechanism. Unlike traditional fusion methods such as simple concatenation or early/late fusion, this approach enables bidirectional interaction, allowing EEG features to selectively attend to audio features and vice versa. This enhances the joint representation and captures dependencies between brain activity and speech.
2. **Patient-centric data splitting:** To avoid data leakage and enhance generalisability, a strict patient-level split is applied—ensuring that no data from the same individual appears across training, validation, or test sets. EEG and audio data from each participant are processed concurrently, preserving inter-modal correlations. Unlike prior studies that apply random splits and risk data leakage, our approach ensures subject-level independence across training, validation, and test sets.

### Overview

This study proposes a multimodal framework for detecting MDD using EEG and audio data. As illustrated in Figure 1, the methodology consists of five sequential stages. Initially, both EEG and audio data undergo preprocessing to ensure signal quality and consistency. The preprocessed EEG signals are transformed into effective connectivity representations-specifically gPDC and direct Directed Transfer Function (dDTF) images—while audio recordings are converted into LBMS images.

These representations serve as inputs to two parallel convolutional encoders: one dedicated to EEG-derived connectivity images and the other to audio spectrograms. Each encoder is augmented with a Convolutional Block Attention Module (CBAM), which enhances salient spatial and channel-wise features by selectively emphasising informative elements within each modality.

Subsequently, a cross-modal attention mechanism fuses the outputs from both encoders. This component enables bidirectional interaction between modalities, allowing EEG features to attend to relevant audio features and vice versa, thereby capturing complementary interdependencies. The resulting joint representation is then passed through fully connected layers to classify instances as MDD or non-MDD, based on the integrated feature space.

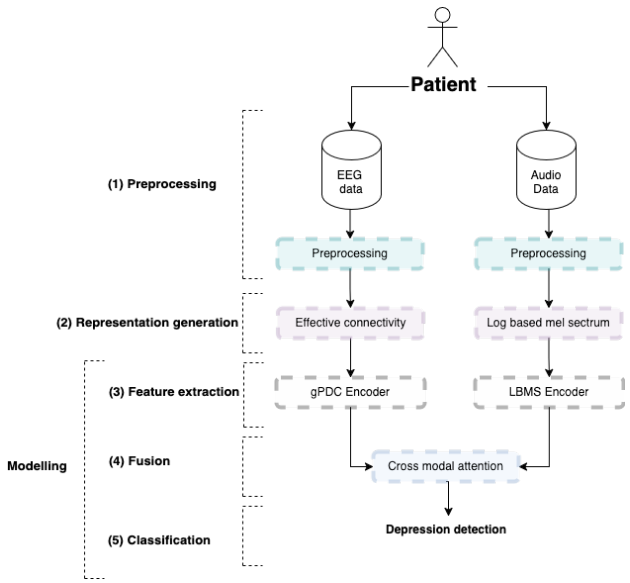


**Figure 1:** Overview of the proposed multimodal framework for MDD detection. The pipeline includes: (1) Preprocessing of EEG and audio data; (2) Generation of gPDC/dDTF and LBMS representations; (3) Feature extraction using modality-specific encoders with CBAM; (4) Cross-modal attention-based fusion; and (5) Classification using fully connected layers.

### Processing of EEG data

EEG signals are inherently susceptible to noise, artefacts, and physiological interference, necessitating rigorous preprocessing to reveal meaningful neural dynamics. From the original 128-channel recordings, 16 electrodes (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, and T6) were selected based on their established relevance in depression detection [3] (Wang et al. 2022), offering a balance between computational efficiency and diagnostic utility.

The raw signals were sampled at 250 Hz and undergo ensemble normalisation to standardise amplitude variations across subjects. Each data point is centred and scaled using the overall mean and standard deviation, ensuring a uniform distribution. Subsequently, re-referencing is performed using an average reference approach to reduce common-mode noise, including physiological and environmental interference.

To further enhance signal fidelity, artefact removal is carried out using Clean Raw Data and Artifact Subspace Reconstruction (ASR), as implemented in the SIFT toolbox (Mullen 2010). ASR effectively suppresses non-neural activity by detecting and at-

tenuating high-variance segments through subspace projection. This is followed by the exclusion of noisy channels and transient artefacts, preserving the neural signal while removing contaminating elements.

A bandpass filter (1–50 Hz) is then applied to focus on neural oscillations within biologically relevant frequency bands. High-pass filtering removes slow drifts, while low-pass filtering reduces high-frequency noise. The filtered signals are subsequently decomposed into canonical EEG frequency bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–80 Hz), facilitating a detailed spectral analysis.

Following decomposition, directional interactions between brain regions are estimated using effective connectivity metrics within each frequency band. These measures enable the construction of frequency-specific connectivity matrices, which are encoded as 2D images that capture both the strength and direction of inter-regional influences. This representation enriches the dataset with spectral and causal information relevant to the diagnosis of MDD, reflecting the established utility of effective connectivity in depression research (Saeedi et al. 2021).

**Effective connectivity**

Effective connectivity refers to the directed influence that one neural population exerts over another. In this study, it is computed from preprocessed EEG signals using a time-varying multivariate autoregressive (tv-MVAR) model. The tv-MVAR framework captures the temporal dynamics of EEG channel interactions, providing the basis for estimating frequency-domain measures that characterise directional brain connectivity.

Among various model-based approaches to effective connectivity, Granger Causality (GC) is widely adopted due to its data-driven formulation. This study employs a parametric GC approach grounded in the tv-MVAR framework, which is well-suited for handling the non-stationary nature of EEG signals.

The tv-MVAR model is formulated as:

$$\mathbf{x}(t) = \sum_{p=1}^{P} \mathbf{A}_p(t)\mathbf{x}(t-p) + \mathbf{e}(t),$$

where $\mathbf{x}(t)$ denotes the EEG signal vector at time $t$, $\mathbf{A}_p(t)$ are the time-varying coefficient matrices for lag $p$, and $\mathbf{e}(t)$ represents the residual noise.

To estimate these coefficients adaptively, the Recursive Least Squares (RLS) algorithm is employed. At each time step, the predicted signal is given by:

$$\hat{\mathbf{x}}(t) = \sum_{p=1}^{P} \mathbf{A}_p(t)\mathbf{x}(t-p),$$

with the prediction error defined as: $\boldsymbol{\epsilon}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$.

The Kalman gain $\mathbf{G}(t)$, which governs the update of the coefficient estimates, is calculated as:

$$\mathbf{G}(t) = \mathbf{P}(t-1)\mathbf{x}(t)^T(\mathbf{x}(t)\mathbf{x}(t)^T + \lambda\mathbf{I})^{-1},$$

and the coefficients are iteratively updated according to:

$$\mathbf{A}_p(t) = \mathbf{A}_p(t-1) + \mathbf{G}(t)\boldsymbol{\epsilon}(t)\mathbf{x}(t-p)^T.$$

To extract frequency-resolved effective connectivity from the tv-MVAR model, this study employs two complementary measures: gPDC and the dDTF. The gPDC quantifies directed interactions by examining how well one signal predicts another in the frequency domain, while accounting for the influence of all other signals. This is achieved by transforming the MVAR model into the frequency domain using the Fourier transform. The spectral matrix of the residuals $\mathbf{R}(f) = \mathbf{e}(f)\mathbf{e}(f)^H$ and the spectral matrix of the data $\mathbf{S}(f) = \mathbf{x}(f)\mathbf{x}(f)^H$ are then used to express the connectivity matrix. The gPDC from node $j$ to node $i$ at frequency $f$ is given by:

$$gPDC_{ij}(f) = \frac{\sigma_{ii}(f) - \sum_{k\neq i} |h_{ij,k}(f)|^2}{\sigma_{ii}(f)},$$

where $\sigma_{ii}(f)$ is the auto-spectral density of node $i$, and $h_{ij,k}(f)$ is the cross-spectral density between nodes $j$ and $i$, conditioned on all nodes except $k$. This formulation captures the direct, frequency-specific causal effect of one EEG channel on another.

The dDTF, in contrast, is derived from the transfer function of the frequency-domain MVAR representation. The process begins with computing the Partial Directed Coherence (PDC) from the transfer matrix $\mathbf{A}(f)$, which is obtained via the Fourier transform of the MVAR coefficients. The Directed Transfer Function (DTF) is then calculated to normalise directional influences across all source nodes. Finally, the dDTF refines this estimate by suppressing indirect effects, providing a clearer measure of the direct causal influence from node $j$ to node $i$ at frequency $f$:

$$\mathrm{dDTF}_{ij}(f) = \frac{DTF_{ij}(f)}{\sum_{k\neq i} DTF_{ik}(f)}.$$

Together, gPDC and dDTF generate matrices encoding the dynamic, directional connectivity between EEG channels. These are visualised as images and serve as input for the subsequent classification of MDD and non-MDD subjects.

**Processing of audio data**

The audio data undergo a structured preprocessing pipeline designed to ensure uniformity and transform raw recordings into LBMS images suitable for machine learning tasks. Recordings are captured at a sampling rate of 44.1 kHz with 24-bit resolution, preserving the full human auditory range (20 Hz to 20 kHz). Each signal is amplitude-normalised to 0 dBFS, thereby mitigating variability introduced by differing recording setups.

To preserve temporal dynamics, recordings—each lasting up to 25 seconds—are segmented into non-overlapping 5-second intervals. These segments are then processed to generate LBMS images by extracting spectral representations from short-time frames using standard signal processing techniques.

**Log-based mel spectrogram (LBMS) extraction**

Each 5-second segment is partitioned into overlapping 20 ms frames, based on the quasi-stationarity assumption that speech signals exhibit local temporal stability. Overlapping frames are used to preserve continuity in the temporal structure. Each frame is then multiplied by a Hamming window to minimise spectral leakage and accentuate central frequency components. The Hamming window is defined as:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1,$$

which tapers the signal, reducing edge effects prior to frequency analysis.

The windowed frames are transformed into the frequency domain via the Short-Time Fourier Transform (STFT). For a signal frame $\mathbf{x}_n$ and its corresponding window $\mathbf{w}_n$, the STFT is given by:

$$\mathbf{Z}(h, k) = \sum_{n=0}^{N-1} \mathbf{x}_{n+h} \cdot \mathbf{w}_n\, e^{-i2\pi\frac{kn}{N}},$$

where $h$ and $k$ represent the time and frequency indices, respectively. This transformation enables the detection of both transient and sustained spectral components.

To model human auditory perception more accurately, the resulting power spectrum is filtered through a Mel filterbank consisting of triangular band-pass filters distributed on the Mel scale. The Mel scale, defined as:

$$\text{Mel}(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right),$$

mimics the ear's sensitivity to frequency—exhibiting a linear resolution below 1 kHz and a logarithmic resolution above 1 kHz. Filterbank energies are aggregated to produce the Mel spectrogram, where each band captures energy within perceptually meaningful frequency regions.

Finally, logarithmic compression is applied to the Mel spectrogram to enhance subtle spectral features while suppressing the dominance of high-energy components. This non-linear transformation yields the LBMS image, offering a compact and perceptually relevant representation of acoustic features that serves as the input to the downstream classification model.

### Model architecture

The proposed framework adopts a dual-stream multimodal architecture designed to learn both modality-specific and cross-modal representations for the effective detection of MDD. As illustrated in Figure 2, the model processes two input modalities in parallel: gPDC images derived from EEG signals and LBMS images derived from audio recordings.

Each modality is fed into a dedicated convolutional encoder, both of which are augmented with a CBAM. These modules refine the feature maps by applying channel and spatial attention mechanisms, thereby enhancing the salience of discriminative features within each modality.

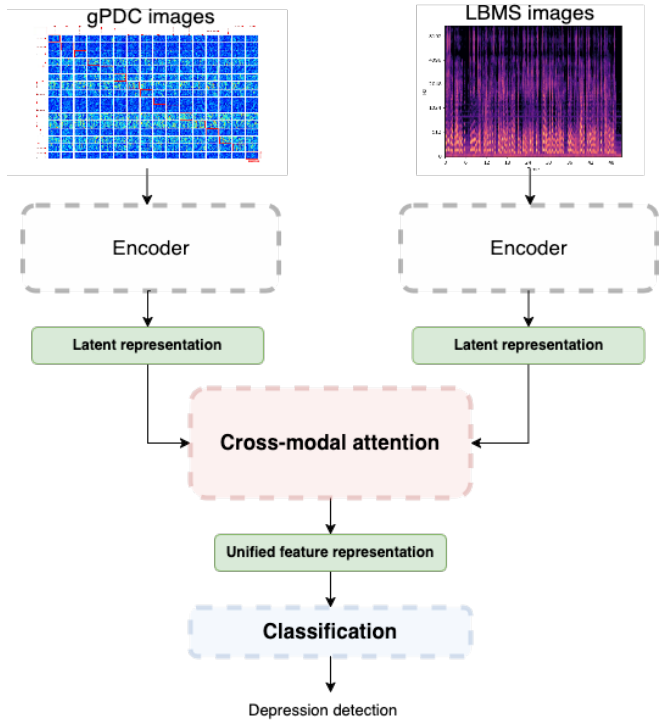


**Figure 2:** Architecture of the proposed multimodal model for MDD detection. EEG-derived gPDC images and audio-derived LBMS images are processed by parallel encoders with CBAM. A cross-modal attention mechanism fuses the extracted features, followed by classification layers to predict MDD.

### Encoders

The encoder architecture, shown in Figure 3, transforms each modality into a compact and expressive feature representation through a series of convolutional processing blocks. Each block comprises a 2D convolutional layer (kernel size: 3X3, activation: ReLU), followed by batch normalisation and a 2D max pooling operation (pool size: 2X2). This structure progressively reduces the spatial resolution while retaining essential features.

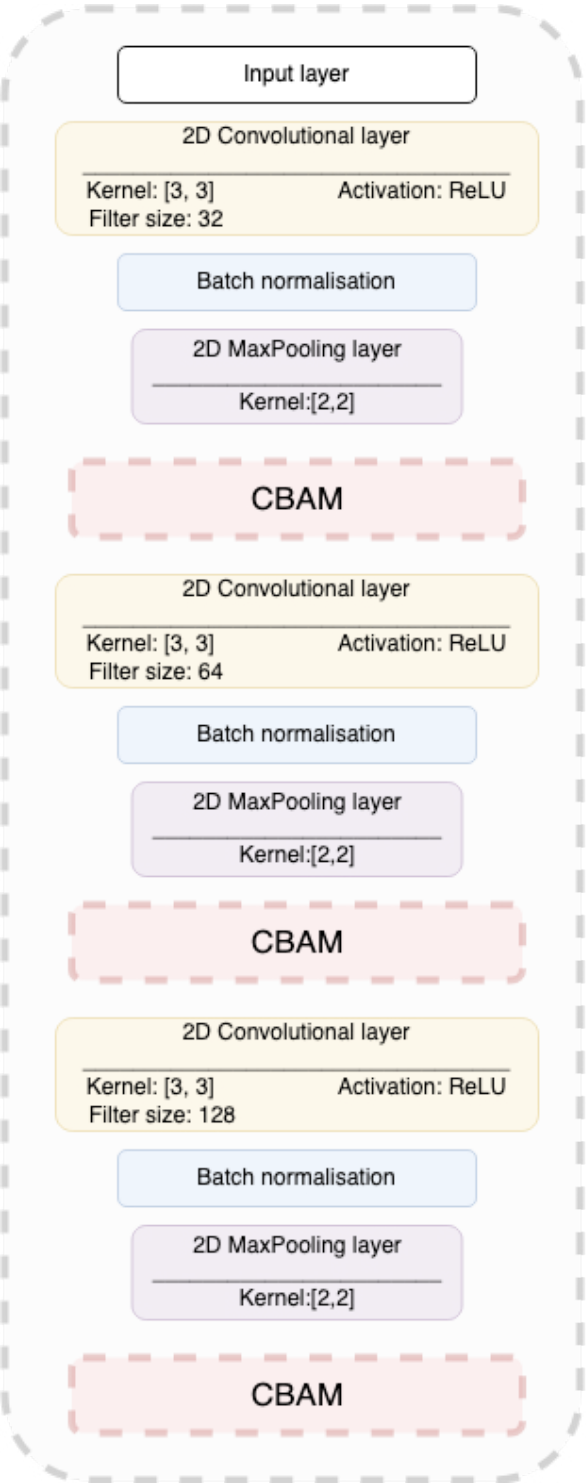


**Figure 3:** Illustration of the proposed encoder architecture. The model begins with an input layer followed by three consecutive convolutional blocks. Each block consists of a 2D convolutional layer (kernel size: 3×3, activation: ReLU), batch normalisation, and a 2D max-pooling layer (pool size: 2×2) to progressively reduce spatial dimensions. After each convolutional block, a Convolutional Block Attention Module is applied to enhance feature representation by focusing on relevant spatial and channel information.

A CBAM is applied after each convolutional block to enhance the discriminative capability of the learned features. In the EEG encoder, CBAM modules selectively amplify patterns associated with diagnostically relevant neural connectivity. In contrast, the audio encoder's CBAMs prioritise frequency and temporal features indicative of speech abnormalities linked to MDD.

By utilising separate encoders for EEG and audio inputs, the architecture retains the unique signal characteristics inherent to each modality—namely, brain connectivity patterns in EEG and spectral-temporal structures in audio.

**Convolutional block attention module**

The CBAM enhances the representational capacity of convolutional neural networks by sequentially applying channel and spatial attention mechanisms to intermediate feature maps. Initially popularised by (Woo et al. 2018), CBAM builds upon earlier foundational work, including that of [9], who explored the integration of spatial and channel attention to improve feature learning.

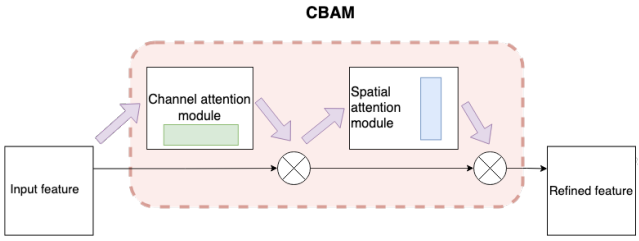


**Figure 4:** The architecture of the Convolutional Block Attention Module, consisting of channel and spatial attention modules. Input features are refined through both modules to enhance relevant information.

As illustrated in Figure 4, CBAM comprises two sequential submodules. The Channel Attention Module (CAM) evaluates the significance of each feature channel, selectively enhancing those that capture discriminative information. The Spatial Attention Module (SAM), on the other hand, focuses on identifying the most informative spatial locations across the feature maps. This sequential attention mechanism enables the network to effectively learn both what to attend to (via channels) and where to attend to (spatial positions), which is particularly valuable in multimodal learning scenarios.

**Channel attention module (CAM)**

The CAM extends the concept of the Squeeze-and-Excitation (SE) block proposed by (Hu et al. 2018), which recalibrates channel-wise responses using global context. CAM improves upon this by combining both global average pooling and max pooling operations, producing two descriptors of shape *C X* 1 *X* 1 that encapsulate complementary channel statistics.

These descriptors are passed through a shared Multi-Layer Perceptron (MLP) consistin of a hidden layer with a reduced dimensionality by a factor of , which helps reduce computational complexity. The two outputs are then merged via element-wise summation and passed through a sigmoid activation function to produce the final channel attention map. This map is multiplied element-wise with the input feature map, enhancing semantically meaningful channels. The fusion of average and max pooling enables CAM to capture both dominant features and global contextual trends.

**Spatial attention module (SAM)**

While CAM focuses on what features to enhance, SAM addresses where the salient information is located. It computes spatial descriptors by applying average and max pooling operations along the channel axis, generating two separates $1 \times H \times W$ maps. These maps are then concatenated and convolved using a 7X7 kernel, followed by a sigmoid activation to generate the final spatial attention map.

This soft attention mask is applied to the feature map through element-wise multiplication, allowing the model to dynamically emphasise spatial regions that are most relevant to the task. The large convolution kernel ensures a broad receptive field, capturing contextual information across the spatial domain.

The sequential application of CAM and SAM in CBAM allows the model to refine its internal feature representations effectively, focusing attention on both the most informative channels and spatial locations. Importantly, this attention refinement is achieved with minimal additional computational cost. CBAM has proven especially beneficial in multimodal architectures, where distinct modalities may contribute complementary information distributed across different dimensions (Zhu et al. 2022).

The latent representations enhanced by CBAM in each encoder are subsequently passed into the cross-modal attention mechanism, where inter-modal relationships are learned and leveraged for the final classification of MDD.

**Cross-modal attention mechanism**

Cross-modal attention facilitates interaction between heterogeneous modalities by allowing features from one modality to inform and refine representations in another. In the proposed model, latent representations produced by the EEG and audio encoders are transformed into query (Q), key (**K**), and value (**V**) matrices, which are then processed by a cross-modal attention mechanism designed to align and integrate complementary information across modalities.

This integration is achieved through scaled dot-product attention, wherein the query is multiplied by the transpose of the key, scaled by the inverse square root of the key dimension, and normalised using a softmax function to produce attention weights. These weights determine the influence of each value vector, allowing the model to generate context-sensitive outputs that reflect inter-modal dependencies.

To capture a broad range of relationships—such as temporal synchrony or spectral complementarity—multi-head attention is employed. Each head operates in a separate subspace, learning distinct patterns of correlation between EEG and audio data.

**Multi-head attention**

Input features are linearly projected into multiple subspaces to form head-specific **Q**, **K**, and **V** matrices. Each head independently applies scaled dot-product attention, producing representations that capture different inter-modal cues—such as correlations between frequency bands in audio and brain connectivity patterns in EEG.

The outputs of all attention heads are concatenated and passed through a final linear layer, yielding a unified cross-modal feature map. This enables targeted, context-aware interaction between EEG and audio representations, enriching the model's understanding of complex multimodal cues associated with MDD. A visual representation of multi-head attention is depicted in Figure 5.

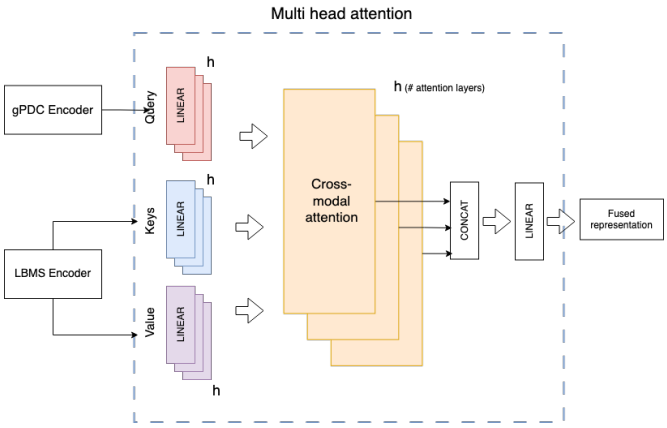


**Figure 5:** Illustration of the multi-head attention mechanism. Inputs from both modalities are linearly projected to generate query, key, and value matrices. These are processed by the cross-modal attention module to integrate modality-specific information.

### Scaled dot-product attention

At the heart of multi-head attention is scaled dot-product attention, which computes the similarity between queries and keys to guide the aggregation of values. Specifically, attention scores are calculated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V},$$

where $d_k$ is the dimension of the key vectors. Scaling by $\sqrt{d_k}$ mitigates the effect of large dot-product magnitudes, stabilising gradients during training. This attention mechanism operates in parallel across all heads, producing refined, context-aware representations that encapsulate key inter-modal relationships.

### Implementation

The cross-modal attention mechanism is implemented in two sequential stages, as shown in Figure 6, enabling bidirectional interaction between the EEG and audio modalities. This design allows each modality to guide the other's feature refinement, fostering deeper alignment and mutual enhancement of salient cues.

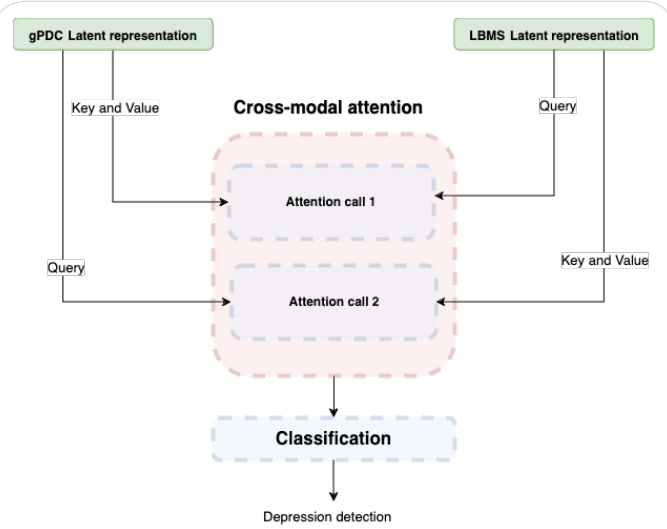


**Figure 6:** Schematic of the cross-modal attention mechanism integrating gPDC and LBMS latent representations. Two attention operations are applied: one using gPDC as the query and LBMS as key/value, and another using LBMS as the query and gPDC as key/value.

In the first stage, features extracted from EEG (gPDC images) are used to form the query matrix $\mathbf{Q}$, while the key $\mathbf{K}$ and value $\mathbf{V}$ matrices are derived from audio features (LBMS images):

$$\text{Attention}_{Stage\ 1} = softmax\left(\frac{\mathbf{Q}_{EEG}\mathbf{K}_{Audio}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{Audio}.$$

This allows EEG features to attend to relevant auditory cues, modelling how neural activity may synchronise with or respond to acoustic patterns associated with depressive states.

In the second stage, the direction of attention is reversed. Audio-derived features are used to form the query, while EEG features serve as the key and value:

$$\text{Attention}_{Stage\ 2} = softmax\left(\frac{\mathbf{Q}_{Audio}\mathbf{K}_{EEG}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{EEG}.$$

This reciprocal mechanism enables the audio modality to selectively focus on relevant EEG patterns, uncovering mutual dependencies that may not be evident when modalities are processed independently.

The outputs of both stages are fused to produce a joint latent representation that retains the most discriminative and complementary features from each modality. This enriched representation is then passed to the classification head for MDD detection.

### Model training

The proposed deep learning architecture was trained using a systematic procedure that incorporated $k$-fold cross-validation, Bayesian optimisation for hyperparameter tuning, and data augmentation to improve generalisability and ensure reliable evaluation.

### Data collection

This study employs raw EEG and audio data from the Multimodal Open Dataset for Mental Disorder Analysis (MODMA) [6], which includes full-brain, 128-channel resting-state EEG recordings and corresponding audio samples. The dataset comprises 52 participants: 23 with depression (16 males, 7 females; aged 16–56) and 29 healthy controls (20 males, 9 females; aged 18–55).

Audio was recorded in a soundproof environment (ambient noise less than 60 dB) using Neumann TLM102 microphones and an RME FIREFACE UCX interface at 44.1 kHz and 24-bit resolution, stored as uncompressed WAV files. This study focuses on the reading task, where participants recited *The North Wind and the Sun*, a standardised text in multilingual acoustic research. Each recording was divided into five segments, yielding 260 LBMS images for the audio modality.

EEG data were acquired in a controlled setting with participants seated quietly in a room, monitored remotely to minimise artefacts. Only the resting-state task was used, in which participants sat with eyes closed for five minutes. Recordings were made with a 128-channel HydroCel Geodesic Sensor Net (EGI, USA) using Net Station v4.5.4 at 250 Hz. Signals were referenced to Cz, with electrode impedance kept below 50 kΩ, and cap sizes adjusted individually.

EEG recordings were saved in .mff format and converted to .mat for preprocessing in MATLAB using EEGLAB. Data from electrodes E1–E128 (Cz referenced) were used to compute effective connectivity via gPDC and dDTF across five standard EEG frequency bands. This resulted in five connectivity images per participant per method, totalling 260 images for each measure

and forming the EEG modality.

**Dataset splitting**

The dataset was divided into training/validation and test subsets in an 80:20 ratio. A participant-level split was enforced to ensure that all data from a given individual appeared exclusively in a single subset. The test set remained entirely unseen throughout training and hyperparameter tuning, thereby serving as an unbiased benchmark for final performance assessment.

The training/validation subset was further divided into $k = 5$ folds for cross-validation, such that each fold served once as the validation set while the remaining $k - 1$ folds were used for training. A patient-level split was also applied during this stage, ensuring that all samples from a given participant were confined to a single fold. This maintained participant independence across folds, mitigated the risk of overfitting and information leakage, and ensured that all available data contributed to both training and validation.

**Hyperparameter tuning**

Hyperparameters were optimised using Bayesian optimisation embedded within the cross-validation framework. Each candidate configuration was evaluated based on the average validation performance across the five folds. The search was guided by a Tree-structured Parzen Estimator (TPE) and spanned 20 trials, with early stopping employed to terminate underperforming configurations.

The hyperparameter search space included: learning rate ($10^{-5}$–$10^{-2}$; selected: 0.0013), batch size (4–32; selected: 8), number of attention heads (2–8; selected: 3), key dimensionality (32–128; selected: 36), dropout rate (0.1–0.5; selected: 0.4), kernel size (3,3)–(7,7); selected: 3,3), number of filters (32, 64, 128; selected: 128), L2 regularisation coefficient ($10^{-6}$–$10^{-3}$; selected: 0.000833), and number of training epochs (10–50; selected: 30).

Following hyperparameter selection, the model was retrained on the full 80% training/validation set and evaluated via 5-fold cross-validation. Final performance was then assessed on the held-out 20% test set, which had remained untouched during prior stages.

To further mitigate overfitting, data augmentation was applied independently to the gPDC and LBMS datasets, expanding each from 260 to 1,300 images.

**Data augmentation**

Data augmentation is a widely used strategy in deep learning to improve model generalisability by artificially expanding the training dataset. It introduces controlled variations to the input data, enabling the model to learn invariant features and reducing the risk of overfitting. In this study, augmentation was applied to both EEG and audio modalities to simulate real-world variability and enhance reliability.

For the EEG-based gPDC and dDTF images, several transformations were applied. Gaussian noise and blurring simulated signal perturbations and spatial imprecision, while random changes in brightness and contrast introduced variability linked to physiological or recording conditions. In addition, weak connections were selectively masked to reflect the sparse nature of brain connectivity, and elastic deformations were applied to mimic spatial distortions due to anatomical or equipment-related differences.

The audio modality, represented as LBMS images, was augmented using SpecAugment. This included time and frequency masking, time warping, amplitude perturbations, and localised noise injection—each intended to replicate natural distortions and variability encountered in speech and environmental noise.

All augmentations were implemented dynamically during training. Randomised transformations were applied in real time, conserving memory while increasing sample diversity. This on-the-fly augmentation strategy encouraged the model to learn generalisable patterns, thereby improving its performance on unseen data. Nonetheless, it is recognised that augmentation cannot fully compensate for limited dataset diversity, and its effectiveness remains contingent on the breadth of the original data.

**Integration**

To ensure the integrity of performance evaluation, particular attention was paid to preventing data leakage—a key concern when working with participant-specific data such as EEG and audio. Improper partitioning may result in inflated performance if the model inadvertently encounters data from the same individual across different subsets.

This study therefore adopts a *Patient-Centric* integration strategy, built upon two core principles. First, data from both modalities (gPDC and LBMS images) are processed jointly for each participant to preserve the temporal and multimodal coherence unique to that individual. Second, data are strictly partitioned according to participant identifiers. This ensures that all data belonging to a single individual are confined to one subset—training, ١ lidation, or testing. The same logic was maintained during $k$-fold cross-validation, where folds were constructed using participant IDs to ensure strict separation.

This strategy addresses two key methodological concerns:

- **Prevention of data leakage:** By eliminating participant overlap across subsets, the model is compelled to learn discriminative patterns related to mental state rather than identity-specific artefacts. This enhances the credibility of performance evaluation.
- **Accommodation of inter-individual variability:** Given that MDD manifests heterogeneously, joint processing of multimodal data at the participant level enables consistent intra-individual feature learning and supports generalisation across diverse symptom presentations.

In contrast, conventional random splitting approaches risk dispersing data from the same participant across different subsets. This may lead to artificially inflated results if the model exploits identity-based cues. The Patient-Centric strategy mitigates this risk by enforcing strict partitioning, thereby encouraging the model to focus on diagnostic features that are genuinely indicative of MDD.

While this approach enhances the methodological integrity of the study, it is acknowledged that the relatively small number of participants constrains statistical power. Nonetheless, the methodology prioritises rigorous performance evaluation and supports generalizability to unseen individuals.

### Experimental protocol

To evaluate the proposed deep learning architecture for MDD classification, four experiments were conducted. These experiments progressively examine the effectiveness of individual modalities, a simplified dual-modality baseline, the complete multimodal system, and the contributions of specific architectural components.

The first experiment (Experiment 1) assesses unimodal performance by training three separate models on individual data types: EEG-derived gPDC images, EEG-derived dDTF images, and speech-derived LBMS images. Each model uses a dedicated encoder with a CBAM, but omits the cross-modal attention mechanism due to the absence of modality fusion. This experiment establishes baseline performance levels and evaluates the standalone predictive capacity of each modality.

The second experiment (Experiment 2) introduces a simplified dual-modality configuration. EEG (gPDC) and speech (LBMS) data are processed using basic convolutional encoders, without CBAM. The extracted features are concatenated and passed to a classifier. This setup serves as a baseline for evaluating the impact of multimodal integration in the absence of attention mechanisms.

The third experiment (Experiment 3) evaluates the full proposed architecture. gPDC and LBMS images are processed using separate encoders, each enhanced with CBAM. Their outputs are then fused using a cross-modal attention mechanism before final classification. This experiment tests the effectiveness of the complete architecture and quantifies the benefit of integrating EEG and speech data using attention-based fusion.

The fourth experiment (Experiment 4) is an ablation study designed to isolate the contributions of CBAM and the cross-modal attention module. Two architectural variants are examined: one excludes CBAM but retains cross-modal attention; the other retains CBAM while replacing cross-modal attention with simple concatenation.

## Results

Model performance was evaluated using standard classification metrics: precision, recall, F1-score, and accuracy. These metrics offer a comprehensive assessment of each model's effectiveness in accurately and reliably identifying MDD cases.

### Experiment 1 – Unimodal models

Table 1 presents the performance of the three unimodal models trained separately on speech-derived LBMS images (Audio only), EEG-derived dDTF images (dDTF only), and EEG-derived gPDC images (gPDC only).

**Table 1:** Performance metrics for unimodal models processing LBMS, dDTF, and gPDC data respectively.

| | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **Audio only** | 0.754 | 0.741 | 0.747 | 0.752 |
| **dDTF only** | 0.801 | 0.767 | 0.783 | 0.791 |
| **gPDC only** | 0.825 | 0.788 | 0.806 | 0.814 |

Among the three models, the gPDC-based model achieved the highest scores across all metrics, followed by the dDTF model, while the audio-only model performed the least effectively. These findings indicate that EEG-derived features—particularly gPDC—provide a more reliable foundation for MDD classification than those extracted from speech alone.

Notably, the gPDC model yielded the highest precision (0.825), indicating strong capability in reducing false positives. It also achieved the highest recall (0.788), reflecting heightened sensitivity in detecting true MDD cases. Its superior F1-score (0.806) and overall accuracy (0.814) further demonstrate its balanced and dependable performance.

## Discussion

The superior performance of EEG-based models highlights the diagnostic value of effective connectivity features in capturing the neurophysiological characteristics of MDD. Unlike LBMS features derived from speech, gPDC and dDTF capture directional interactions between brain regions—providing a more direct representation of altered neural dynamics associated with depression.

The consistent outperformance of gPDC over dDTF may be attributed to its frequency-domain normalisation, which mitigates variability in signal power across frequency bands. This enhances its sensitivity to discriminative connectivity patterns. Furthermore, gPDC is well-suited to multivariate data, enabling it to model complex network-level interactions often implicated in MDD pathology. While dDTF also reflects directional connectivity, its relatively limited normalisation and broader signal integration may reduce its sensitivity to subtle neural changes.

Overall, these results affirm the relevance of EEG-based effective connectivity features—particularly gPDC—as a reliable and informative modality for automated MDD detection.

### Experiment 2 – Baseline model

This experiment evaluates the baseline multi-modal model, comparing its performance to the unimodal models trained on either audio or gPDC data alone. Despite employing a simplified fusion strategy—concatenation of features from basic encoders without attention modules—the baseline model consistently outperforms its unimodal counterparts across all evaluation metrics.

The baseline model as shown in Table 2, achieves a precision of 0.876, surpassing both the audio-only model (0.754) and the gPDC-only model (0.825), indicating improved accuracy in identifying true positive MDD cases. Recall improves to 0.869, compared to 0.741 and 0.788 for audio and gPDC respectively, suggesting enhanced sensitivity. The F1-score reaches 0.872, reflecting a more effective balance between precision and recall. Furthermore, the overall accuracy increases to 0.874, outperforming both the audio-only (0.752) and gPDC-only (0.814) models.

**Table 2:** Comparison of performance metrics (precision, recall, F1-score, and accuracy) for the "Audio only" model, "gPDC only" model, and baseline model (Experiment 2).

| 2-5 | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **Audio only** | 0.754 | 0.741 | 0.747 | 0.752 |
| **gPDC only** | 0.825 | 0.788 | 0.806 | 0.814 |
| **Baseline** | 0.876 | 0.869 | 0.872 | 0.874 |

These findings underscore the benefit of multi-modal integration, even when implemented using a basic fusion strategy. The improved performance indicates that EEG and speech features provide complementary information, and their combination enhances the model's capacity to detect MDD more accurately. This result reinforces the value of incorporating multiple

physiological and behavioural modalities for reliable mental health assessment.

### Experiment 3 – Full model

The performance of the Full model during training and validation is illustrated in Figure 7, showing consistent learning behaviour across all five cross-validation folds. Training accuracy steadily increases and approaches 1.0 in each fold, while validation accuracy rapidly improves and stabilises at similarly high levels. Although minor fluctuations are observed, the validation trends suggest strong generalisation. Correspondingly, both training and validation losses exhibit a sharp initial decline before plateauing, indicating efficient convergence with minimal signs of overfitting.

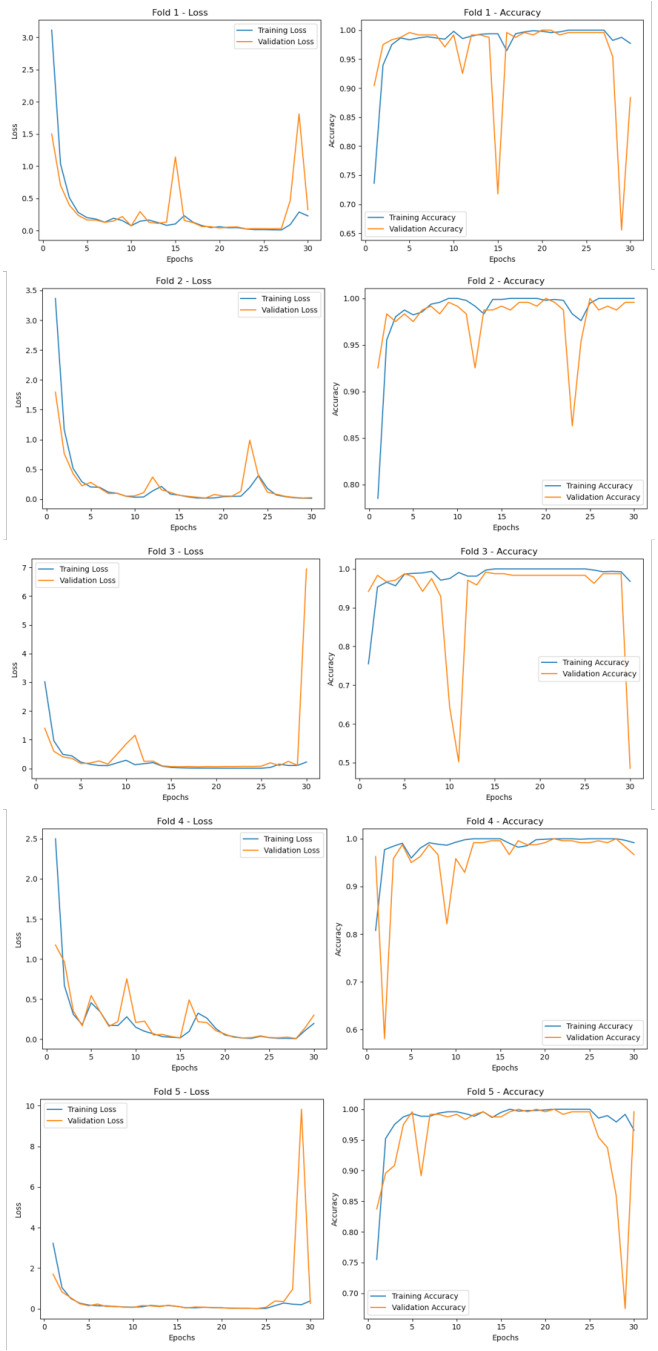


**Figure 7:** Training and validation accuracies and losses across five folds for the Full model over 30 epochs. Each fold shows consistent learning dynamics, with training accuracy approaching 1.0 and validation accuracy stabilising at high levels. Training and validation losses decline rapidly and then stabilise, indicating efficient optimisation and minimal overfitting.

The cross-validation performance metrics, summarised in Table 3, further validate the model's reliability. The Full model achieves a precision of 0.9810±0.0047, recall of 0.9793±0.0201, F1-score of 0.9778±0.0105, and accuracy of 0.9786±0.0114, reflecting highly consistent and balanced predictions across folds. These results affirm the efficacy of integrating gPDC and LBMS data for reliable detection of MDD.

**Table 3:** Cross-validation performance metrics for the Full model across five folds. Standard deviations indicate inter-fold variability.

| Metric | Mean value | Standard deviation |
|---|---|---|
| Precision | 0.9810 | 0.0047 |
| Recall | 0.9793 | 0.0201 |
| F1-Score | 0.9778 | 0.0105 |
| Accuracy | 0.9786 | 0.0114 |

On the held-out test set, the model maintains high classification performance, as illustrated in the confusion matrix (Figure 8). It correctly identifies 97.82% of MDD cases and 97.39% of non-MDD cases, with low false positive (2.61%) and false negative (2.18%) rates. These results confirm the model's strong generalisation and discriminative ability across classes.

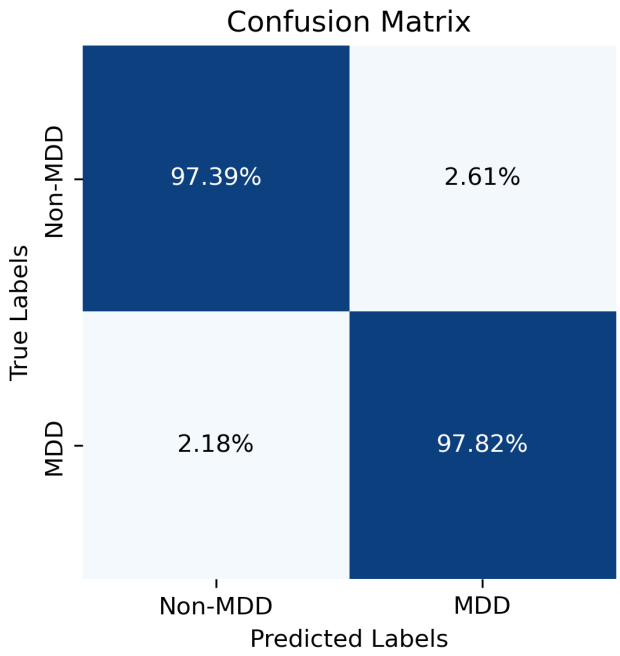


**Figure 8:** Confusion matrix of the Full model expressed as percentages. The model correctly classifies 97.82% of MDD and 97.39% of non-MDD cases, with low false positive (2.61%) and false negative (2.18%) rates.

### Comparison with related systems

To contextualise the performance of the proposed multimodal model, its results are compared against existing systems developed using the MODMA dataset. As shown in Table 4, the model introduced in this study achieves an accuracy of 97.86%, positioning it among the most effective multimodal approaches for detecting MDD.

Notably, while some prior studies report higher accuracy values—exceeding 99% in certain cases—these figures often arise under less stringent evaluation protocols. A substantial number of existing models employ random data splitting (denoted as Method A), whereby training, validation, and test data are

drawn from the same pool of participants. This practice introduces a high risk of data leakage, potentially inflating performance metrics by allowing models to learn participant-specific characteristics.

For instance, Zheng et al. (2023) and Gupta et al. [18] report accuracies of 99.0% and 99.9%, respectively, under random splits. However, such results may not reflect true generalisability. In contrast, the present study adopts a patient-centric strategy (Method B), ensuring strict separation of subjects across all data subsets. This approach mitigates overfitting and offers a more realistic measure of model performance in clinical applications.

Among the few studies that also employ Method B, such as Chen et al. [10], reported performance is considerably lower (86.49%), highlighting the relative strength of the current architecture. The gap between models using different splitting protocols—particularly in studies such as Hu et al. [21] and Ning et al. (2024), which report accuracies of 79.0% and 87.44%, respectively—reinforces the critical importance of rigorous data partitioning when evaluating models intended for clinical deployment.

**Table 4:** Comparison of recent multimodal models for MDD detection using the MODMA dataset. Method A: random split; Method B: patient-centric split. In several prior works, the use of Method A is inferred from the description of the experimental setup rather than explicitly stated.

| Study | Model | Split Method | Accuracy (%) |
|---|---|---|---|
| (Hu et al. 2024) [21] | LLM | A | 79.00 |
| (Chen et al. 2022) [10] | GNN | B | 86.49 |
| (Ning et al. 2024) | SVM | A | 87.44 |
| (Zhang et al. 2022) | Multi-agent system | A | 92.30 |
| (Ahmed et al. 2023) | CNN-LSTM | A | 96.43 |
| (Qayyum et al. 2023) | Transformer | A | 97.31 |
| **This study** | **Encoders** | **B** | **97.86** |
| (Zheng et al. 2023) | Transformer | A | 99.00 |
| (Gupta et al. 2023) | Federated DL | A | 99.90 |

**Experiment 4 – Ablation study**

To assess the individual contributions of the model's core components, an ablation study was performed, comparing the full model against two modified configurations and the baseline model. The full model which integrates CBAM modules within both encoders and employs a cross-modal attention mechanism achieved the highest performance, with precision, recall, F1-score, and accuracy all nearing 97.9%.

In Configuration 1, CBAM modules were retained in each encoder to maintain refined intra-modal representations; however, the cross-modal attention mechanism was removed. Instead, features from the two modalities were simply concatenated before classification. While this configuration still delivered strong results—achieving 93.1% accuracy—it lacked the sophisticated inter-modal feature alignment provided by cross-modal attention.

Configuration 2 adopted the opposite approach, removing CBAM modules while preserving cross-modal attention. This enabled the model to learn inter-modal relationships, but the absence of CBAM reduced the quality of features extracted within each modality. Consequently, performance declined further, with accuracy dropping to 90.6%.

The baseline model, which excluded both CBAM and cross-modal attention, exhibited the lowest performance across all evaluation metrics, reaching only 87.4% accuracy. Without any intra- or inter-modal attention mechanisms, this configuration clearly demonstrates the critical role that these components play in achieving effective multimodal classification.

The results of these experiments are summarized in Table 5.

**Table 5:** Performance comparison of the "Full", baseline models and the two ablation study configurations. The full model integrates both CBAM and cross-modal attention, achieving the highest accuracy and F1-score. Configuration 1 demonstrates the effect of CBAM without cross-modal attention, while Configuration 2 highlights the performance with cross-modal attention but no CBAM. The baseline configuration, lacking both CBAM and cross-modal attention, achieves the lowest performance metrics.

| Model Configuration | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **Full Model** | 0.981 | 0.979 | 0.978 | 0.979 |
| **Configuration 1** | 0.938 | 0.927 | 0.932 | 0.931 |
| **Configuration 2** | 0.915 | 0.904 | 0.909 | 0.906 |
| **Baseline** | 0.876 | 0.869 | 0.872 | 0.874 |

**Discussion:** The ablation study results underscore the complementary benefits of combining CBAM and cross-modal attention in the model architecture. Configuration 1 highlights the importance of CBAM in enhancing intra-modal feature representation by selectively emphasising salient spatial and channel-wise information. However, without an effective mechanism to align and integrate features across modalities, simple concatenation constrains the model's ability to exploit the complementary strengths of the different modalities.

Conversely, Configuration 2 demonstrates that the potential of cross-modal attention to capture rich inter-modal interactions depends heavily on the quality of the input features. In the absence of CBAM, which serves to refine these features within each modality, the benefits of cross-modal attention are limited—resulting in diminished overall performance.

The baseline model, which lacks both mechanisms, reinforces the necessity of incorporating both intra-modal and inter-modal attention. The full model's superior accuracy and balanced evaluation metrics validate the architectural decision to combine CBAM with cross-modal attention, as each mechanism plays a vital and complementary role in extracting and integrating discriminative features from EEG and audio data for reliable Major Depressive Disorder detection.

## Discussion

The experimental findings highlight the value of combining neurophysiological (EEG) and vocal (speech) biomarkers through carefully designed attention mechanisms for the effective classification of MDD. Three main insights emerge from the proposed multimodal framework.

First, the superior performance of gPDC compared to both dDTF and LBMS in unimodal configurations reinforces the discriminative strength of EEG-derived features in identifying MDD. However, the performance gap observed between unimodal and multimodal configurations indicates the limitations of relying on a single modality. The improvement achieved by the baseline multimodal model, through simple

feature concatenation, illustrates the complementary nature of EEG and speech data. While EEG captures changes in neural connectivity, speech reflects behavioural and affective vocal cues—each offering different but mutually supportive perspectives relevant to MDD classification.

Second, the strong performance of the "Full" model is attributed to its ability to model dependencies between modalities. The CBAM modules improve intra-modal representation by identifying important spatial and channel-level patterns within each data type. At the same time, the cross-modal attention mechanism captures relationships across modalities by computing attention weights that guide how features from each modality influence one another. As shown in the ablation study (Table 5, this dual attention structure provides a significant improvement over models using either attention mechanism alone.

Third, the patient-centric approach adopted in this study addresses a known limitation in previous work. While some studies report very high performance under randomised sample-level splitting—sometimes exceeding 99.9%—the use of participant-level partitioning in this study enables a more clinically relevant assessment. The proposed model achieves an accuracy of 97.86% under these more rigorous conditions, indicating stronger potential for generalisation to real-world scenarios. This is particularly notable when compared with other studies applying similar evaluation protocols, such as [10], which reported an accuracy of 86.49%.

Overall, this study demonstrates that a carefully constructed multimodal architecture can offer an effective and interpretable solution for automated MDD detection. By combining the neural specificity of EEG with the behavioural expressiveness of speech through dual attention mechanisms, the proposed model captures complementary information from both modalities.

A summary of the performance of the respective models is provided in Table 6.

**Software used**

The analysis pipeline incorporated several specialised software tools to support reliable data preprocessing, feature extraction, and model development.

EEGLAB, a widely-used MATLAB-based toolbox, played a central role in EEG preprocessing. Its comprehensive functionality—including event and channel data import, filtering, epoching, and artefact rejection—was enhanced through the use of Independent Component Analysis (ICA), which effectively separates neural signals from noise, thereby improving signal fidelity [12].

Effective connectivity metrics were extracted using the Source Information Flow Toolbox (SIFT). This open-source MATLAB toolkit facilitated advanced model fitting and connectivity estimation, producing gPDC and dDTF measures that visualise directional information flow in the brain. These representations were essential inputs to the deep learning models, and SIFT's analytical tools supported the interpretation of connectivity patterns relevant to MDD (Mullen 2010).

Model implementation and training were conducted using TensorFlow and Keras. These frameworks provided the necessary flexibility and computational efficiency to construct convolutional encoder architectures enhanced with CBAM and cross-modal attention modules.

Hyperparameter tuning was automated using Optuna, which systematically optimised model configurations to improve performance.

## Conclusion

This study proposed and evaluated deep learning methodologies for the diagnosis of MDD, the results demonstrate the effectiveness of combining modalities, with the fusion models significantly outperforming uni-modal counterparts.

Key findings include the superior performance of gPDC over dDTF in capturing relevant EEG connectivity patterns, the benefits of attention mechanisms such as CBAM and cross-modal attention for enhancing feature representation, and the success of a patient-centred data split in preventing data leakage. The best-performing model achieved 97.86% accuracy, highlighting the diagnostic potential of multi-modal learning in mental health applications.

While promising, the study is limited by dataset size and demographic diversity, as well as the computational demands of attention-based fusion. Additionally, the binary classification focus restricts broader clinical applicability.

Future work should explore additional modalities, larger and more diverse datasets, longitudinal designs, and personalised modelling approaches. Collaboration with clinical experts is essential to improve model interpretability and alignment with the neurobiological and behavioral underpinnings of MDD.

## References


1. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H, Subha DP. Automated EEG-based screening of depression using deep convolutional neural network. Comput Methods Programs Biomed. 2018; 161: 103–113.
2. Ahmed S, Yousuf MA, Monowar MM, Hamid A, Alassafi MO. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. IEEE Access. 2023; 11: 99847–99861.
3. Akar SA, Kara S, Agambayev S, Bilgiç V. Nonlinear analysis of EEGs of patients with major depression during different emotional states. Comput Biol Med. 2015; 67: 49–60.
4. Ay B, Yildirim O, Talo M, et al. Automated depression detection using deep representation and sequence learning with EEG signals. J Med Syst. 2019; 43: 1–12.
5. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories-IA and -II in psychiatric outpatients. J Pers Assess. 1996; 67: 588–597.
6. Cai H, Gao Y, Sun S, et al. MODMA dataset: A multi-modal open dataset for mental-disorder analysis. arXiv. 2020; arXiv:2002.09283.
7. Campisi P, La Rocca D. Brain waves for automatic biometric-based user recognition. IEEE Trans Inf Forensic Secur. 2014; 9: 782–800.
8. Cannizzaro M, Harel B, Reilly N, Chappell P, Snyder PJ. Voice acoustical measurement of the severity of major depression. Brain Cogn. 2004; 56: 30–35.
9. Chen L, Zhang H, Xiao J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc IEEE Conf Comput Vis Pattern Recognit. 2017. p. 5659–5667.

10. Chen T, Hong R, Guo Y, Hao S, Hu B. MS²-GNN: Exploring GNN-based multimodal fusion network for depression detection. IEEE Trans Cybern. 2023; 53: 7749–7759.

11. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015; 71: 10–49.

12. Delorme A, Makeig S. EEGLAB: An open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004; 134: 9–21.

13. Demertzis K, Rantos K, Magafas L, Iliadis L. A cross-modal dynamic attention neural architecture to detect anomalies in data streams. Appl Sci. 2023; 13: 9648.

14. Dubagunta SP, Vlasenko B, Magimai-Doss M. Learning voice source related information for depression detection. In: Proc IEEE Int Conf Acoust Speech Signal Process. 2019. p. 6525–6529.

15. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA. 2017; 317: 1517.

16. Gao Y, Cao Z, Liu J, Zhang J. A novel dynamic brain network in arousal for brain states and emotion analysis. Math Biosci Eng. 2021; 18: 7440–7463.

17. Goldman LS, Nielsen NH, Champion HC. Awareness, diagnosis, and treatment of depression. J Gen Intern Med. 1999; 14: 569–580.

18. Gupta C, Khullar V, Goyal N, et al. Cross-silo privacy-preserving federated multimodal system for identification of major depressive disorder. Diagnostics. 2024; 14: 43.

19. He L, Chan JCW, Wang Z. Automatic depression recognition using CNN with attention mechanism from videos. Neurocomputing. 2021; 422: 165–175.

20. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc IEEE Conf Comput Vis Pattern Recognit. 2018. p. 7132–7141.

21. Hu Y, Zhang S, Dang T, et al. Exploring large-scale language models to evaluate EEG-based multimodal data for mental health. In: Proc ACM Int Joint Conf Pervasive Ubiquitous Comput. 2024. p. 412–417.

22. Huang Z, Epps J, Joachim D, Sethu V. NLP methods for acoustic and landmark event-based features in speech-based depression detection. IEEE J Sel Top Signal Process. 2019; 14: 435–448.

23. Jiang H, Hu B, Liu Z, et al. Detecting depression using ensemble logistic regression based on speech features. Comput Math Methods Med. 2018; 2018: 1–9.

24. Jurysta F, Kempenaers C, Lancini J, et al. Altered interaction between cardiac vagal influence and delta sleep EEG in major depressive disorder. Acta Psychiatr Scand. 2010; 121: 236–239.

25. Kang M, Kwon H, Park JH, Kang S, Lee Y. Deep-asymmetry matrix image for depression pre-screening. Sensors. 2020; 20: 6526.

26. Koller-Schlaud K, Ströhle A, Bärwolf E, Behr J, Rentzsch J. EEG frontal asymmetry and theta power in depression. J Affect Disord. 2020; 276: 501–510.

27. Kraepelin E. Manic-depressive insanity and paranoia. Edinburgh: E & S Livingstone; 1921.

28. Liu W, Zhang C, Wang X, et al. Functional connectivity of major depressive disorder using ongoing EEG. Clin Neurophysiol. 2020; 131: 2413–2422.

29. Mallol-Ragolta A, Zhao Z, Stappen L, Cummins N, Schuller B. Hierarchical attention networks for depression detection. In: Proc Interspeech. 2019.

30. Zung WWK. A self-rating depression scale. Arch Gen Psychiatry. 1965; 12: 63–70.