



ARTICLE

A Multi-Input Neural Network for Early Detection of Neurological Disorders from Vocal and Sleep Data

Md. Shujan Shak¹, Nabila Rahman², Fuad Mahmud³, Ashim Chandra Das¹, M. F. Mridha⁴ and Md. Jakir Hossen^{5,*}

¹Doctor of Computer Science Program, University of the Potomac, Virginia, VA, USA

²Doctor of Information Technology Program, Trine University, Detroit, MI, USA

³Information Assurance and Cybersecurity, Gannon University, Erie, PA, USA

⁴Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh

⁵Center for Advanced Analytics (CAA), COE for Artificial Intelligence, Faculty of Engineering & Technology (FET), Multimedia University, Melaka, Malaysia

*Corresponding Author: Md. Jakir Hossen. Email: jakir.hossen@mmu.edu.my

Received: 23 February 2026; Accepted: 20 May 2026; Published: 22 October 2026

ABSTRACT: Early detection of neurological disorders is critical for effective treatment planning and improved quality of life. This study proposes a multi-input deep neural network that integrates vocal biomarkers and clinical sleep-related features to improve diagnostic accuracy. The model processes each modality through separate neural branches before combining high-level representations for final classification. We evaluate the approach using two public datasets: a Parkinson's disease dataset containing 1195 voice samples and a sleep-disorder dataset with 80 patient records. Experimental results show that the proposed model outperforms classical machine learning baselines and single-modality deep learning models, achieving an accuracy of 92.5%, precision of 90.2%, recall of 94.1%, F1-score of 92.1%, and an AUC-ROC of 0.953. The architecture demonstrates consistent improvements over early and late fusion strategies, demonstrating the benefit of modality-specific representation learning. This work provides a scalable and non-invasive diagnostic framework with strong potential for clinical screening and monitoring.

KEYWORDS: Neurological disorders; deep learning; multimodal learning; Parkinson's disease; clinical decision support

1 Introduction

Neurological disorders such as Parkinson's disease and sleep-related cognitive decline affect millions globally and represent a major source of disability and mortality [1–3]. Early diagnosis is crucial for initiating timely intervention, improving treatment outcomes, and improving patient quality of life [4]. However, conventional diagnostic procedures typically require extensive clinical testing, neurological expertise, and invasive techniques, which can be costly, inaccessible, and less effective for early-stage detection.

Recent advances in artificial intelligence (AI), particularly deep learning, have opened new possibilities for non-invasive and data-driven approaches to medical diagnostics. Among various modalities, vocal biomarkers and sleep-related physiological measurements have emerged as promising indicators of neurological health. Changes in speech patterns—such as jitter, shimmer, and pitch irregularities—are early manifestations of neurodegeneration, while altered sleep dynamics and breathing patterns are associated with cognitive decline and neurological comorbidities. Singh and Dash [5] reviewed the use of machine

learning and deep learning methods for early diagnosis of neurological disorders, particularly Parkinson's and Alzheimer's diseases. Their study highlights the limitations of traditional diagnostic practices, which often miss early signs due to subtle symptoms or patient unawareness. The authors emphasize that while conventional Machine Learning (ML) approaches rely heavily on manual feature engineering, deep learning models offer improved accuracy by learning features directly from data. The chapter also discusses recent advances and introduces a smartphone-based monitoring system for Alzheimer's patients using sensor data, demonstrating the potential of AI-driven, non-invasive healthcare solutions.

Nunes et al. [6] introduced a hybrid diagnostic framework that combines a rule-based expert system with multicriteria decision-making methods to support early identification of psychological disorders. It leverages structured knowledge from the DSM-5 and integrates it with the MACBETH method to analyze qualitative control events related to disorders such as schizophrenia, bipolar disorder, depression, anxiety, and more. This model allows for comparative evaluation of symptoms, helping clinicians derive consistent diagnostic insights even when data is incomplete or subjective. The system uses expert-defined rules and confidence factors to generate diagnostic suggestions, offering a structured, explainable, and adaptable approach to psychological assessment. The proposed method emphasizes flexibility, decision support, and effective detection, particularly in professional environments where mental health issues are often underdiagnosed.

Despite growing interest in multimodal learning for medical applications, existing approaches often rely on early feature concatenation or single-stream models, which limit the model's ability to preserve modality-specific patterns [7]. Such methods may overlook the distinct statistical distributions and the diagnostic relevance of each data type [8]. Moreover, there is a lack of architectures that are both interpretable and efficient enough for deployment in real-world screening settings. To address these challenges, we propose a multi-input deep neural network that independently processes vocal and clinical sleep features before high-level fusion. This design allows the model to learn rich, non-linear representations within each modality and leverage their combined predictive power for classification. We evaluate the model using two publicly available datasets: one comprising 1195 voice recordings for Parkinson's disease diagnosis, and another with 80 polysomnographic sleep records capturing key physiological metrics.

This work introduces a multimodal learning framework that differs from conventional approaches in several key aspects. Recent studies in multimodal healthcare learning highlight that most existing systems rely on direct fusion strategies or assume aligned multimodal data, which may fail to preserve modality-specific characteristics and limit generalization in heterogeneous clinical settings [9,10]. In contrast, the proposed architecture is designed to operate on independent datasets without requiring matched samples. The model learns modality-specific representations through separate network branches and performs integration at the representation level, enabling effective knowledge fusion across heterogeneous data sources. Furthermore, the use of a lightweight architecture tailored to tabular clinical data improves generalization under limited data conditions. These characteristics distinguish the proposed framework from existing multimodal learning approaches and demonstrate its applicability in real-world clinical scenarios where aligned multimodal datasets are not always available. Our contributions are summarized as follows:

- We design a dual-branch multi-input neural network tailored to heterogeneous data, combining speech-derived acoustic features with sleep-related clinical measurements.
- We evaluate the model across multiple metrics—accuracy, precision, recall, F1-score, and AUC-ROC—showing that it outperforms traditional baselines and single-modality deep learning models.
- We conduct ablation studies to analyze the contribution of each modality and architectural component, demonstrating the effectiveness of modality-specific processing and late fusion.

- We demonstrate the model's applicability to real-world screening systems through its non-invasive input sources, scalability, and diagnostic performance.

The remainder of this paper is organized as follows: [Section 2](#) reviews related work. [Section 3](#) presents the methodology, including the model architecture and training strategy. [Section 4](#) reports the experimental results. [Section 5](#) discusses the findings, limitations, and future directions. [Section 6](#) concludes the paper.

2 Related Work

Recent advances in deep learning have accelerated the development of diagnostic models across various domains in healthcare [11]. In the context of neurological disorder detection, prior research has largely focused on unimodal inputs, typically leveraging either acoustic signals or physiological data independently [12,13]. While these approaches have demonstrated potential in controlled settings, they often lack the robustness required for generalizable, real-world application.

Voice-based models have gained attention due to their non-invasive nature and ease of data collection [14]. Many studies have explored the use of handcrafted acoustic features or end-to-end neural architectures for detecting early signs of neurodegenerative conditions such as Parkinson's disease [15]. Islam et al. [15] provide an in-depth review of machine learning and deep learning techniques applied to the diagnosis of Parkinson's disease, with a specific focus on voice and handwriting data. The authors present a broad comparison of algorithms, datasets, and features across more than 60 studies, and evaluate their effectiveness in classification tasks. They also highlight the progression from traditional classifiers to advanced deep neural networks, and discuss the potential of emerging biomarkers in enhancing diagnostic precision. The paper also summarizes publicly available datasets, feature extraction methods, and classification performance using metrics such as accuracy and recall. Overall, the study serves as a guide for future research by identifying key trends, technical challenges, and opportunities for improving early detection of Parkinson's disease through automated, data-driven approaches. However, these models typically operate in isolation and do not account for other physiological indicators that may correlate with disease onset or progression.

In parallel, sleep data—particularly signals derived from polysomnography—has been used to assess sleep disorders and associated neurological decline [16–18]. Approaches using statistical modeling, classical machine learning, and recurrent neural networks have been proposed, focusing on features such as oxygen desaturation, apnoea-hypopnoea indices, and Electroencephalography (EEG) patterns [19]. Despite their effectiveness in isolated tasks, these models are constrained by the assumption that a single modality is sufficient for accurate diagnosis. Hireš et al. [20] introduce a novel ensemble approach using convolutional neural networks (CNNs) for the automatic detection of Parkinson's disease (PD) from vowel-based voice recordings. The authors propose a multiple fine-tuning (MFT) method to bridge the gap between pretrained source tasks and the target PD classification task, improving feature transfer. By using spectrograms of sustained vowels and applying Gaussian blurring, the input data is adapted for deep CNN training. The ensemble model combines predictions from CNNs fine-tuned on different intermediate datasets, which improves classification robustness. Evaluated on the PC-GITA dataset, the model achieves up to 99% accuracy for certain vowels, outperforming existing techniques and demonstrating the feasibility of accurate, non-invasive PD diagnosis using short voice samples.

Multimodal learning has emerged as a promising direction that integrates diverse data sources such as speech, clinical metrics, and sensor signals [21]. However, much of the existing work employs early fusion strategies, where features are concatenated at the input level before feeding them into a shared network. While computationally simple, this design often fails to capture modality-specific representations and can degrade performance due to feature scale disparity or noise. Cheng et al. [22] propose a novel deep learning framework, called MML-DMS, for the simultaneous classification of sleep stages and sleep disorders

using multimodal physiological signals, including EEG, Electrocardiography (ECG), and Electromyography (EMG). The system employs multiple convolutional neural networks (CNNs) to independently analyze each modality for both tasks, and then fuses their outputs through shallow neural networks to make the final decision. Evaluated on the PhysioNet CAP Sleep Database, the method achieves high accuracy—94.34% for sleep stage recognition and 99.09% for sleep disorder detection—outperforming previous models. The study demonstrates that combining both multimodal data and multilabel outputs leads to superior diagnostic performance compared to traditional single-modality or single-task approaches, offering an effective solution for automating complex sleep diagnostics.

Few studies have investigated architectures that maintain independent processing pipelines for heterogeneous clinical data before performing late or intermediate fusion [23]. There is a lack of systematic exploration into how such architectures compare against standard baselines in terms of diagnostic performance and interpretability. Limited attention has also been given to evaluating these models on publicly available datasets that combine both vocal and clinical data, leaving a gap in benchmarking and reproducibility. The need for architectures that support modality-aware representation learning, enable interpretability, and remain practical for deployment in healthcare settings is increasingly recognized. However, challenges remain in effectively combining signals from distinct domains, avoiding overfitting in small clinical datasets, and maintaining generalizability across patient populations. This motivates the development of models that integrate diverse sources of non-invasive data in a structured and adaptive way—an objective addressed in this study.

3 Methodology

This section presents the architecture, training procedure, and evaluation framework of the proposed multi-input deep neural network for early detection of neurological disorders using vocal and sleep data. The model learns modality-specific representations from voice biomarkers and clinical measurements and integrates them at a higher level through a unified deep learning pipeline.

3.1 Data Preprocessing

To ensure robust training and generalization, we implemented advanced preprocessing techniques for both the vocal dataset (Parkinson's Disease) and the clinical sleep dataset (Sleep Disorder). This includes normalization, encoding, imputation, feature transformation, and data splitting for effective deep learning training.

3.1.1 Preprocessing of Vocal Features (Parkinson's Dataset)

The Parkinson's dataset contains $n = 1195$ voice recordings with binary labels indicating the presence or absence of Parkinson's disease. Let $\mathbf{X}^{(v)} \in \mathbb{R}^{n \times d_v}$ denote the acoustic feature matrix, where $d_v = 23$.

The vocal features were obtained from the publicly available Parkinson's dataset in pre-extracted tabular form, rather than computed from raw audio. The feature set consists of established acoustic biomarkers used in neurological voice analysis, including measures related to fundamental frequency, jitter, shimmer, harmonicity, and nonlinear vocal signal variation. Accordingly, this study uses the dataset-provided vocal descriptors as input features and applies normalization and outlier control before model training.

Z-Score Normalization

Each feature was standardized to have zero mean and unit variance to ensure numerical stability during training.

$$\hat{x}_{ij}^{(v)} = \frac{x_{ij}^{(v)} - \mu_j}{\sigma_j} \quad (1)$$

where $x_{ij}^{(v)}$ is the j th feature of the i th sample, and μ_j, σ_j are the mean and standard deviation of feature j across all samples.

Outlier Clipping

All standardized values were clipped to the interval $[-3, 3]$ to mitigate the impact of outliers:

$$\hat{x}_{ij}^{(v)} = \min(3, \max(-3, \hat{x}_{ij}^{(v)})) \quad (2)$$

3.1.2 Preprocessing of Clinical Features (Clinical Sleep Dataset)

The clinical sleep dataset contains $n = 80$ samples with demographic and physiological attributes relevant to sleep-related disorders. Denote the input matrix as $\mathbf{X}^{(s)} \in \mathbb{R}^{n \times d_s}$, where $d_s = 13$ after processing.

The sleep-related features were obtained from the public clinical sleep dataset in structured tabular form and include demographic and physiological variables commonly used in sleep assessment. These variables include blood pressure measurements, sex, and clinical indices such as ODI, AHI, and AI. Since the dataset provides these attributes in pre-compiled form, this work focuses on feature engineering, missing-value imputation, categorical encoding, and normalization rather than raw polysomnographic signal extraction.

Blood Pressure Feature Engineering

We separated the compound attribute $\text{BP}_{\text{sys}}/\text{BP}_{\text{dia}}$ into systolic and diastolic components:

$$\text{BP}_{\text{sys}}, \text{BP}_{\text{dia}} = \text{Split}(\text{BP}_{\text{sys}/\text{dia}}, \varepsilon/\varepsilon) \quad (3)$$

Missing Value Imputation

We applied k -Nearest Neighbors imputation for missing values in attributes such as oxygen desaturation index (ODI), apnea-hypopnea index (AHI), and apnea index (AI), which are standard clinical indicators used to assess sleep-related breathing disorders:

$$\hat{x}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} x_j \quad (4)$$

where $\mathcal{N}_k(i)$ denotes the $k = 5$ nearest neighbors of sample i .

Categorical Encoding

The binary categorical variable sex was converted to numerical format using label encoding:

$$\text{sex}_{\text{encoded}} = \begin{cases} 0, & \text{if Female} \\ 1, & \text{if Male} \end{cases} \quad (5)$$

Min-Max Normalization

All continuous features were normalized to the $[0, 1]$ range:

$$\hat{x}_{ij}^{(s)} = \frac{x_{ij}^{(s)} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (6)$$

3.1.3 Label Definition and Harmonization

Each dataset retains its original binary label definition corresponding to its respective condition. For the vocal dataset, labels indicate the presence or absence of Parkinson's disease. For the clinical sleep dataset, labels are defined based on clinical severity criteria (e.g., AHI thresholds).

To maintain a unified learning objective, both datasets are mapped to a binary classification task that represents the presence or absence of disorder-related conditions. This formulation aligns the prediction objective across modalities without requiring subject-level correspondence between samples.

From a clinical perspective, both Parkinson's disease and sleep-related disorders are associated with neurological dysfunction and often exhibit overlapping physiological and behavioral indicators. Vocal impairments are well-documented early markers of Parkinson's disease, while sleep disturbances such as abnormal AHI patterns are linked to neurological and neurodegenerative conditions. Therefore, the proposed framework is designed as a generalized neurological screening model that leverages complementary indicators from different clinical domains. Importantly, this formulation does not merge the two conditions into a single diagnosis but aligns them under a shared screening objective, enabling the model to learn modality-specific patterns relevant to neurological health.

3.1.4 Dataset Merging and Splitting

The vocal dataset and the clinical sleep dataset used in this study originate from independent cohorts and do not contain matched samples corresponding to the same individuals. Therefore, direct subject-level multimodal feature concatenation is not feasible.

Instead of constructing combined feature vectors at the sample level, the proposed framework processes each dataset independently through modality-specific input branches. The model learns separate latent representations for vocal and clinical sleep features, which are subsequently integrated at a higher level within the network during the classification stage. This approach enables the extraction of complementary diagnostic patterns from heterogeneous data sources without requiring one-to-one correspondence between samples.

Accordingly, no data synthesis or artificial pairing was performed, and both datasets were used in their original form within the unified learning framework.

Each dataset was partitioned into training, validation, and testing subsets using a stratified split:

$$\mathcal{D}_{\text{train}} = 70\%, \quad \mathcal{D}_{\text{val}} = 15\%, \quad \mathcal{D}_{\text{test}} = 15\% \quad (7)$$

Stratified sampling was applied to maintain class balance across all subsets.

3.1.5 Batch Preparation

All subsets were converted into PyTorch tensors and loaded in mini-batches of size $B = 32$ during training:

$$\mathcal{B} = \left\{ (\mathbf{X}_b, \mathbf{y}_b) \right\}_{b=1}^{\lceil N/B \rceil} \quad (8)$$

This preprocessing pipeline ensured data quality and consistency across modalities, supporting effective training of the proposed multi-input neural network.

3.2 Proposed Methodology

This section describes the proposed model, including its design and key architectural details. Fig. 1 illustrates a high-level view of the overall model structure.

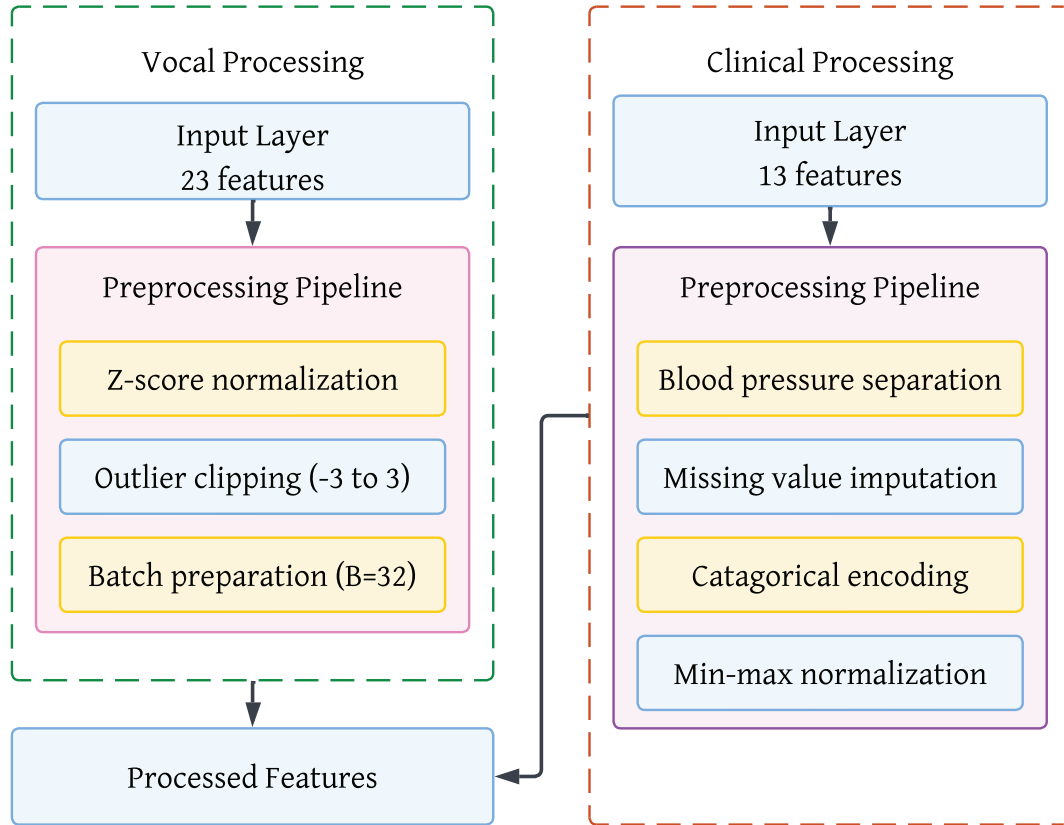


Figure 1: Modality-specific data preprocessing pipeline for vocal and clinical inputs, illustrating normalization, feature engineering, and transformation steps applied independently before downstream integration.

3.2.1 Model Overview

Let $\mathbf{x}^{(v)} \in \mathbb{R}^{d_v}$ denote acoustic feature vectors from the vocal dataset, and $\mathbf{x}^{(s)} \in \mathbb{R}^{d_s}$ denote clinical feature vectors from the sleep dataset. Since the two datasets originate from independent cohorts, the model is designed to learn modality-specific representations instead of paired sample-level mappings.

The objective is to learn a function:

$$f : \mathbb{R}^{d_v} \cup \mathbb{R}^{d_s} \rightarrow \{0, 1\} \quad (9)$$

where the network processes each modality through separate branches and integrates their learned representations at a higher level to perform classification.

3.2.2 Model Architecture

We designed a dual-branch deep neural network with modality-specific input layers followed by a fusion mechanism as shown in Fig. 2. The architecture is a compact dual-branch network with clearly defined layers in each modality-specific branch, followed by a fusion and classification stage. To improve clarity, the structural configuration, including layer sizes, activation functions, and training settings, is summarized in Tables 1 and 2, allowing a concise overview without relying solely on mathematical expressions.

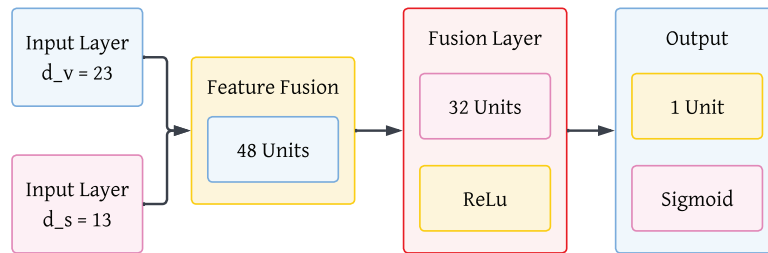


Figure 2: Detailed architectural diagram of the multi-input neural network, showing the dual-branch design, layer configurations, and feature fusion mechanism.

Table 1: Layer-wise architecture of the proposed multi-input neural network.

Branch	Layer Type	Output Shape	# Parameters	Activation
Voice Branch	Input Layer	(23,)	0	–
	Dense Layer 1	(64,)	1536	ReLU
	Dense Layer 2	(32,)	2080	ReLU
Sleep Branch	Input Layer	(13,)	0	–
	Dense Layer 1	(32,)	448	ReLU
	Dense Layer 2	(16,)	528	ReLU
Fusion + Classifier	Concatenation	(48,)	0	–
	Dense Layer (Fusion)	(32,)	1568	ReLU
	Output Layer	(1,)	33	Sigmoid
Total Trainable Parameters				6193

Table 2: Summary of model hyperparameters and training configuration.

Component	Configuration
Input Dimensions	Voice: 23, Sleep: 13
Voice Branch Layers	Dense (64) → Dense (32), ReLU activation
Sleep Branch Layers	Dense (32) → Dense (16), ReLU activation
Fusion Layer	Concatenation (48) → Dense (32), ReLU
Output Layer	Dense (1), Sigmoid activation
Loss Function	Binary Cross-Entropy
Optimizer	Adam ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Batch Size	32

(Continued)

Table 2 (continued)

Component	Configuration
Epochs	100
Dropout Rate	0.5 (applied after dense layers)
Weight Regularization	L_2 with $\lambda = 10^{-4}$
Early Stopping	Patience = 10 epochs (validation loss)
Initialization	Xavier uniform initialization
Framework	PyTorch (Python 3.10)

The mathematical formulation below directly corresponds to the layer-wise operations defined in the architecture tables. Each equation represents a standard feedforward transformation within the respective network branch or fusion layer.

The voice branch processes $\mathbf{x}^{(v)}$ using fully connected layers with non-linear activations:

$$\mathbf{h}_1^{(v)} = \text{ReLU}(\mathbf{W}_1^{(v)} \mathbf{x}^{(v)} + \mathbf{b}_1^{(v)}) \quad (10)$$

$$\mathbf{h}_2^{(v)} = \text{ReLU}(\mathbf{W}_2^{(v)} \mathbf{h}_1^{(v)} + \mathbf{b}_2^{(v)}) \quad (11)$$

The clinical features are processed similarly:

$$\mathbf{h}_1^{(s)} = \text{ReLU}(\mathbf{W}_1^{(s)} \mathbf{x}^{(s)} + \mathbf{b}_1^{(s)}) \quad (12)$$

$$\mathbf{h}_2^{(s)} = \text{ReLU}(\mathbf{W}_2^{(s)} \mathbf{h}_1^{(s)} + \mathbf{b}_2^{(s)}) \quad (13)$$

The outputs of the two branches are concatenated and passed through shared dense layers for joint representation learning:

$$\mathbf{h}_{\text{fusion}} = [\mathbf{h}_2^{(v)} \parallel \mathbf{h}_2^{(s)}] \quad (14)$$

$$\mathbf{z}_1 = \text{ReLU}(\mathbf{W}_f \mathbf{h}_{\text{fusion}} + \mathbf{b}_f) \quad (15)$$

$$\hat{y} = \sigma(\mathbf{w}_o^\top \mathbf{z}_1 + b_o) \quad (16)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, yielding the predicted probability of neurological disorder.

3.2.3 Architectural Details

The proposed model consists of two independent input branches for processing vocal and sleep-related features. Each branch includes two dense layers followed by a feature fusion stage and a final classification head. Table 1 presents the full architecture of the network. The network was carefully designed to balance capacity and generalization. The voice branch handles higher-dimensional input and thus uses wider initial layers. The clinical branch receives lower-dimensional input and uses a narrower configuration. The final fused feature vector (48-dimensional) is projected into a compact decision space using a shared dense layer before binary classification. Dropout with a probability of $p = 0.5$ was applied after each dense layer during training. All layers were initialized using Xavier uniform initialization to stabilize gradients.

For a fair comparison with temporal architectures, a multimodal Bidirectional Long Short-Term Memory (BiLSTM) baseline with late fusion was implemented. In this model, the clinical sleep features are processed using a BiLSTM layer to capture sequential dependencies, while the vocal features are processed using a dense neural network due to their tabular representation. The outputs from both branches are concatenated and passed through fully connected layers for final classification.

3.3 Training and Implementation Details

This subsection summarizes the training configuration, optimization strategy, and regularization techniques used in the proposed model. To improve clarity and reproducibility, the key architectural and training hyperparameters are consolidated in [Table 2](#).

3.3.1 Loss Function

We employed binary cross-entropy loss to quantify the classification error:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \quad (17)$$

This loss was minimized using the Adam optimizer with backpropagation.

3.3.2 Regularization

To prevent overfitting, we applied dropout with a probability $p = 0.5$ after each dense layer and included L_2 weight regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \sum_l \|\mathbf{W}_l\|_2^2 \quad (18)$$

where λ is a hyperparameter set to 10^{-4} .

3.3.3 Optimization and Training Strategy

The model was trained using the Adam optimizer with a learning rate of $\alpha = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, for $E = 100$ epochs with a mini-batch size of $B = 32$. Early stopping based on validation loss with a patience of 10 epochs was applied to prevent overfitting. All experiments were implemented in PyTorch using Python 3.10 and executed on a system equipped with an NVIDIA GPU (12 GB VRAM). Hyperparameters were selected through grid search based on validation performance.

The relatively small size of the clinical sleep dataset ($n = 80$) introduces a risk of overfitting. To mitigate this, several strategies were employed. The model architecture was intentionally kept lightweight to avoid excessive model capacity. Regularization techniques including dropout and L_2 weight penalties were applied, and early stopping based on validation loss was used to prevent over-training. In addition, a 5-fold cross-validation strategy (described in [Section 3.3.4](#)) was adopted to ensure robustness. This allows the model to be evaluated across multiple data splits rather than relying on a single partition. The consistent performance observed across folds indicates that the model generalizes well despite the limited dataset size. All experiments were conducted with a fixed random seed. Random seed initialization was applied across all relevant libraries, including NumPy and PyTorch, to ensure consistent data splitting, weight initialization, and training behavior across runs.

3.3.4 Evaluation Protocol

The evaluation protocol was designed to ensure reliable and reproducible assessment of the proposed model. Each dataset was independently partitioned into training, validation, and testing subsets using a stratified split of 70%, 15%, and 15%, respectively. Model training was performed exclusively on the training set, while hyperparameter tuning and early stopping were guided by validation set performance. Final evaluation metrics were computed on the held-out test set to provide an unbiased estimate of model performance.

Stratified sampling was employed during dataset splitting to maintain proportional representation of both classes across all subsets. Model performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, providing a comprehensive assessment of classification performance. As the vocal and sleep datasets originate from independent cohorts, subject-level pairing is not present. Therefore, data separation was performed at the sample level within each modality, ensuring that no overlap occurs between training, validation, and test sets. This prevents data leakage and maintains the integrity of the evaluation process.

K-fold cross-validation was employed as a primary validation strategy to improve evaluation robustness and address potential overfitting associated with limited dataset size. A 5-fold cross-validation strategy was adopted, where the dataset was divided into five equal subsets. In each iteration, four folds were used for training and one fold was used for validation/testing, and this process was repeated five times so that each fold served as the evaluation set once. The final performance metrics were computed by averaging the results across all folds. This strategy yields a more stable estimate of model generalization and reduces variance compared to a single train-test split. The inclusion of k-fold cross-validation ensures that the reported performance, including the 92.5% accuracy, is reproducible across different data splits and is not dependent on a single partition of the dataset. The dataset splitting process was performed using a fixed random seed to ensure consistent partitioning across experimental runs.

4 Results Analysis

This section presents a comprehensive evaluation of the proposed multi-input neural network using the two datasets. We assess performance using multiple classification metrics, compare the model with baselines, evaluate per-modality contributions, and study the effect of architectural and training choices.

4.1 Dataset Description

This study utilizes two publicly available datasets containing vocal and sleep-related clinical features for neurological disorder detection.

The first dataset is a Parkinson's disease voice dataset containing 1195 voice recordings represented by 23 acoustic features. These features include fundamental frequency measures, jitter, shimmer, harmonicity, and nonlinear descriptors such as Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), spread measures, and Pitch Period Entropy (PPE). Each sample is labeled to indicate the presence or absence of Parkinson's disease. This dataset serves as the vocal modality in the proposed framework and is publicly available at: <https://www.kaggle.com/datasets/ucimachinelearning/parkinsons-voice-dataset>.

The second dataset is a clinical sleep dataset consisting of 80 samples with demographic and physiological attributes relevant to sleep-related disorders. The dataset includes variables such as blood pressure, sex, and clinical indices including apnea-hypopnea index (AHI), oxygen desaturation index (ODI), and apnea index (AI). Feature preprocessing includes blood pressure separation, missing-value imputation, and normalization. This dataset is publicly available at: <https://www.kaggle.com/datasets/yfrite/polysom/data>.

Both datasets originate from independent cohorts and are not subject-matched. Therefore, the proposed framework performs modality-specific feature learning followed by high-level fusion rather than subject-level multimodal integration.

4.2 Experimental Results

4.2.1 Performance Metrics

Fig. 3 summarizes the classification performance of the proposed model using accuracy, precision, recall, F1-score, and AUC-ROC. The model achieves an accuracy of 0.925, indicating strong overall predictive performance. The precision score of 0.902 shows that 90.2% of predicted positive cases are correctly identified, while the recall of 0.941 shows that 94.1% of actual positive cases are detected. This high recall is particularly important in early disease detection, where missing true disorder cases can have serious clinical consequences. The F1-score of 0.921 confirms a strong balance between precision and recall, and the AUC-ROC of 0.953 demonstrates excellent discriminative capability across classification thresholds. Together, these results indicate that the proposed model provides reliable and balanced classification performance.

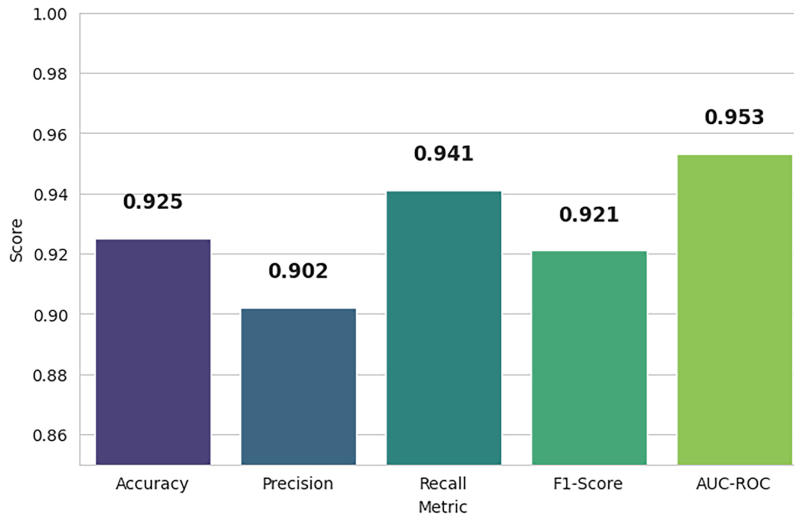


Figure 3: Performance metrics of the proposed model.

4.2.2 Baseline Comparison

The results in **Table 3** demonstrate the effectiveness of the proposed multi-input neural network compared with classical, deep learning, and multimodal baselines. Among the classical models, Support Vector Machine (SVM) achieves the strongest performance, with 0.861 accuracy and 0.901 AUC-ROC, outperforming Logistic Regression and Random Forest. The single-modality deep learning models, including Multilayer Perceptron (MLP), CNN, and BiLSTM, further improve performance, highlighting the benefit of nonlinear feature learning. In particular, the BiLSTM applied to the sleep dataset achieves strong recall, highlighting the relevance of temporal patterns in clinical sleep features.

To ensure a fair comparison with temporal architectures, a multimodal BiLSTM baseline with late fusion was implemented. This model achieves competitive performance, with 0.914 accuracy, 0.890 precision, 0.932 recall, 0.911 F1-score, and 0.946 AUC-ROC. These results demonstrate the effectiveness of incorporating temporal modeling within a multimodal framework.

Table 3: Comparison with baseline models. (DNN: Deep Neural Network).

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.812	0.795	0.837	0.815	0.860
Random Forest	0.845	0.832	0.857	0.844	0.889
SVM (RBF kernel)	0.861	0.851	0.871	0.861	0.901
MLP (Single Input)	0.883	0.874	0.892	0.883	0.924
CNN (1D, Vocal Only)	0.876	0.867	0.881	0.874	0.918
BiLSTM (Sleep Only)	0.868	0.856	0.881	0.868	0.913
Early Fusion DNN	0.902	0.891	0.913	0.902	0.937
Multimodal BiLSTM (Late Fusion)	0.914	0.890	0.932	0.911	0.946
Proposed Multi-Input DNN	0.925	0.902	0.941	0.921	0.953

Early fusion improves performance further, reaching 0.902 accuracy and 0.937 AUC-ROC. However, the proposed multi-input neural network achieves the best results across all reported metrics, with 0.925 accuracy, 0.902 precision, 0.941 recall, 0.921 F1-score, and 0.953 AUC-ROC. These results show that modality-specific processing followed by high-level fusion yields more robust and accurate predictions than both unimodal models and multimodal temporal baselines.

4.2.3 Cross-Validation Analysis

A 5-fold cross-validation strategy was applied to assess the stability and generalization performance of the proposed model. The dataset was divided into five equal subsets, and in each iteration, four folds were used for training while the remaining fold was used for evaluation. This process was repeated five times so that each subset served as the evaluation set once.

Table 4 presents the performance across all folds. The model achieves an average accuracy of 0.925, with precision, recall, and F1-score of 0.902, 0.941, and 0.921, respectively, and an AUC-ROC of 0.953. The variation across folds remains minimal, with accuracy ranging from 0.918 to 0.931 and AUC-ROC from 0.949 to 0.956, indicating stable performance across different data splits. These results confirm that the model maintains strong predictive capability and does not rely on a specific partition of the dataset, demonstrating robustness against overfitting despite the relatively limited dataset size.

Table 4: 5-fold cross-validation performance of the proposed model.

Fold	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Fold 1	0.918	0.895	0.936	0.915	0.949
Fold 2	0.927	0.904	0.942	0.923	0.954
Fold 3	0.931	0.908	0.945	0.926	0.956
Fold 4	0.922	0.899	0.940	0.919	0.951
Fold 5	0.928	0.905	0.943	0.924	0.955
Average	0.925	0.902	0.941	0.921	0.953

4.2.4 Model Interpretability Analysis

Model interpretability was analyzed using SHAP (SHapley Additive exPlanations) values to identify the features that most strongly influenced the predictions of the proposed model. This analysis provides a transparent view of how vocal and clinical sleep variables contribute to the classification outcome. Fig. 4 presents the SHAP summary plot. The analysis indicates that vocal perturbation features such as jitter- and shimmer-related measures, along with clinical sleep indicators including AHI, ODI, and blood pressure variables, contribute strongly to the model's predictions. The observed distribution of SHAP values suggests that the model relies on clinically meaningful patterns across both modalities, thereby improving interpretability and supporting its relevance for clinical decision support. These results provide clinically meaningful insights by identifying key vocal and sleep-related biomarkers influencing the model's predictions, thereby enhancing transparency and trust in the proposed diagnostic framework.

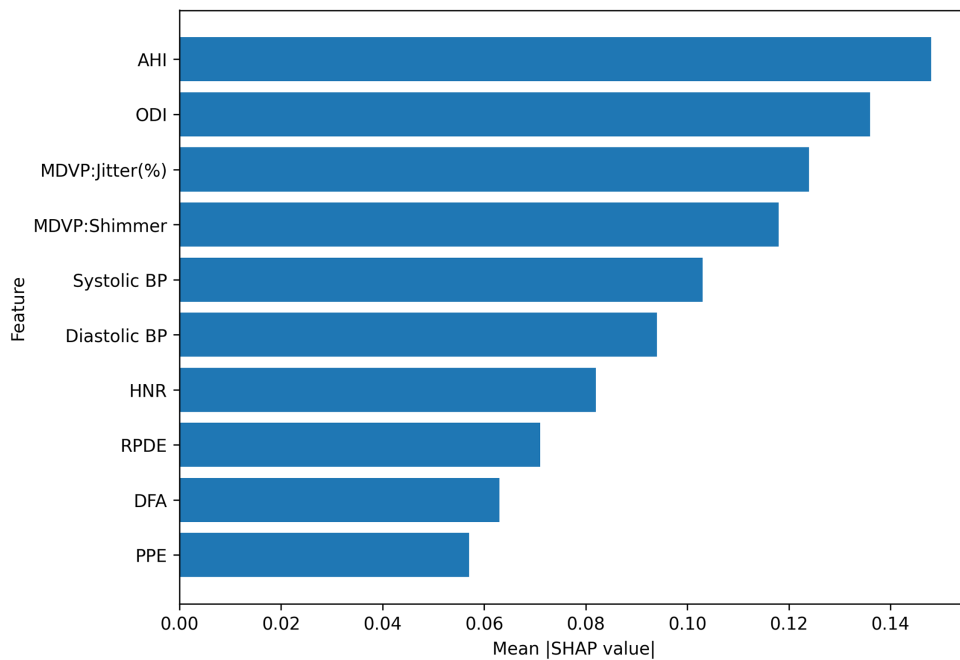


Figure 4: SHAP-based feature importance analysis of the proposed model. Features are ranked according to their contribution to the model output across the evaluation samples.

4.2.5 Per-Modality Performance

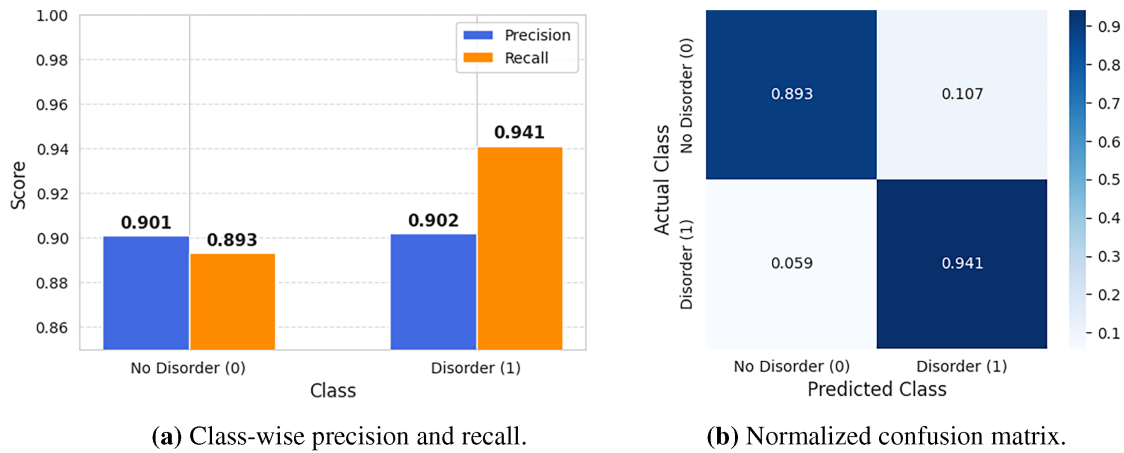
Table 5 presents an ablation study examining the contribution of each input modality. The Voice-Only model achieves 0.891 accuracy, indicating that vocal features alone provide strong discriminative information. The Sleep-Only model attains 0.878 accuracy, confirming that sleep-related clinical features are also informative, although slightly less predictive than vocal features in this setting. Combining both modalities through early fusion increases accuracy to 0.902, showing that multimodal integration improves classification performance. Late fusion further raises accuracy to 0.911, suggesting that preserving modality-specific representations before integration is beneficial. The proposed multi-input model achieves the strongest overall performance, with 0.925 accuracy, 0.902 precision, 0.941 recall, 0.921 F1-score, and 0.953 AUC-ROC. This result confirms that independent branch-wise feature learning followed by high-level fusion provides the most effective approach for the classification task.

Table 5: Modality ablation study.

Feature Input	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Voice Only	0.891	0.882	0.902	0.892	0.927
Sleep Only	0.878	0.867	0.891	0.879	0.914
Voice + Sleep (Early Fusion)	0.902	0.891	0.913	0.902	0.937
Voice + Sleep (Late Fusion)	0.911	0.898	0.926	0.912	0.945
Multi-Input Model (Ours)	0.925	0.902	0.941	0.921	0.953

4.2.6 Class-Wise Performance and Confusion Matrix Analysis

The model demonstrates strong performance across both classes in terms of precision, recall, and overall classification behavior (Fig. 5). For the No Disorder class, it achieves a precision of 0.901, meaning 90.1% of instances predicted as “No Disorder” are correctly classified. The recall of 0.893 indicates that 89.3% of actual “No Disorder” cases are correctly identified. For the Disorder class, the model performs slightly better with a precision of 0.902, indicating that 90.2% of positive predictions are correct. The recall of 0.941 shows that the model successfully identifies 94.1% of actual disorder cases, minimizing false negatives. This higher recall is particularly important for sensitive applications such as medical diagnosis, where missing positive cases can have serious consequences.

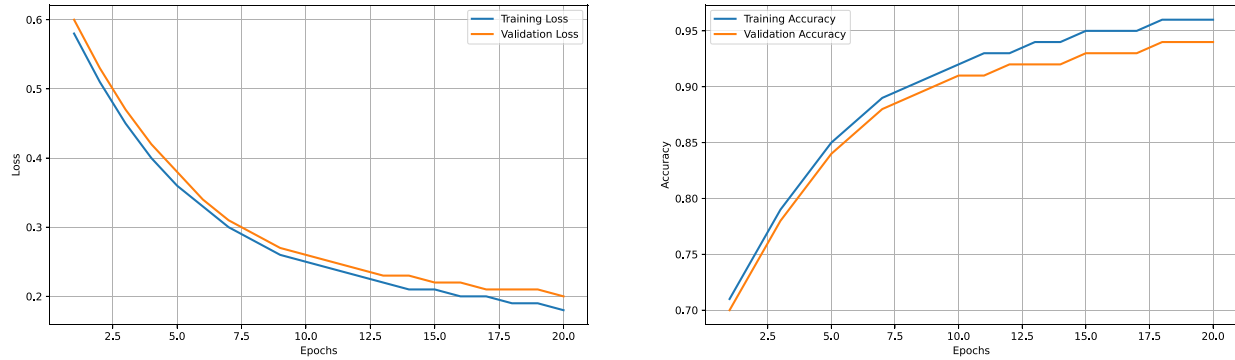
**Figure 5:** Performance evaluation: (a) class-wise precision and recall, and (b) confusion matrix.

The confusion matrix further illustrates these results by showing how actual instances are classified into predicted classes. The model correctly classifies 89.3% of No Disorder cases, while 10.7% are misclassified as Disorder. For the Disorder class, 94.1% of cases are correctly identified, with only 5.9% misclassified as No Disorder.

4.2.7 Training Curve Analysis

To evaluate the learning behavior and generalization capacity of the proposed model, we plotted the training and validation loss and accuracy across training epochs (Fig. 6a,b). As shown in Fig. 6a, both the training and validation loss decrease steadily and converge without significant divergence, indicating that the model does not suffer from overfitting. The validation loss closely tracks the training loss throughout, suggesting good generalization performance. Fig. 6b supports this observation. The training and validation

accuracy increase consistently across epochs, and both curves plateau in the later stages with minimal oscillation. The validation accuracy remains slightly lower than the training accuracy, as expected, but the gap is narrow, showing that the model is neither underfitting nor overfitting.



(a) Training vs. Validation Loss over Epochs.

(b) Training vs. Validation Accuracy over Epochs.

Figure 6: Model training performance: (a) Loss and (b) Accuracy over epochs.

4.2.8 Impact of Dropout Rate

Table 6 evaluates the sensitivity of the proposed model to different dropout rates. Regularization is essential for reducing overfitting, and this experiment identifies the optimal dropout rate for balancing generalization and performance. The dropout rate significantly influences the trade-off between generalization and overfitting. Without dropout (0.0 rate), the model attains 0.901 accuracy, but it may overfit the training data, as suggested by a lower AUC-ROC score of 0.932. Introducing dropout at 0.3 enhances performance, increasing accuracy to 0.911 and AUC-ROC to 0.944, indicating improved generalization. The best performance is observed at 0.5 dropout, achieving 0.925 accuracy, 0.902 precision, 0.941 recall, 0.921 F1-score, and the highest AUC-ROC of 0.953, signifying an optimal balance between model complexity and generalization. However, increasing dropout beyond 0.5 results in a decline in accuracy, with 0.6 dropout yielding 0.912 accuracy, though recall remains relatively high. A 0.7 dropout further reduces accuracy to 0.897, highlighting excessive regularization that limits the model's ability to learn discriminative patterns. These findings suggest that a moderate dropout rate of 0.5 is the most effective, preventing overfitting while maintaining strong classification performance.

Table 6: Effect of dropout rate on model performance.

Dropout Rate	Accuracy	Precision	Recall	F1-Score	AUC-ROC
0.0	0.901	0.886	0.914	0.900	0.932
0.3	0.911	0.897	0.928	0.912	0.944
0.5	0.925	0.902	0.941	0.921	0.953
0.6	0.912	0.893	0.936	0.914	0.947
0.7	0.897	0.872	0.918	0.894	0.929

4.2.9 Effect of Hidden Layer Size

The size of the hidden layer in the fusion component significantly influences the model's ability to learn and generalize (Table 7). A smaller hidden layer (16 neurons) achieves 0.902 accuracy but may limit

representational capacity, as reflected in the 0.934 AUC-ROC score. Increasing the layer size to 32 neurons results in the best performance across all metrics, reaching 0.925 accuracy, 0.902 precision, 0.941 recall, 0.921 F1-score, and 0.953 AUC-ROC, indicating an optimal balance between complexity and generalization. Expanding the hidden layer beyond this size produces diminishing returns. A 64-neuron layer still performs well (0.918 accuracy, 0.946 AUC-ROC) but slightly underperforms the 32-neuron setting. Further increases to 128 neurons and 256 neurons show a gradual decline in performance, with AUC-ROC decreasing to 0.942 and 0.937, respectively, suggesting that excessively large layers introduce unnecessary complexity and potential overfitting. These results indicate that 32 neurons in the fusion layer provide the best trade-off between expressiveness and generalization, ensuring high classification accuracy while limiting overfitting.

Table 7: Effect of fusion layer size on performance.

Fusion Layer Size	Accuracy	Precision	Recall	F1-Score	AUC-ROC
16	0.902	0.887	0.920	0.903	0.934
32	0.925	0.902	0.941	0.921	0.953
64	0.918	0.893	0.936	0.914	0.946
128	0.913	0.888	0.931	0.909	0.942
256	0.905	0.877	0.925	0.900	0.937

5 Discussion

The proposed multi-input neural network demonstrates strong potential for early detection of neurological disorders by integrating heterogeneous data sources—namely vocal biomarkers and clinical sleep-related features. Unlike conventional approaches that rely on either a single modality or naive feature concatenation, our architecture preserves the unique statistical properties of each modality through dedicated input branches. This design enables the model to learn modality-specific patterns before high-level feature fusion, contributing to the observed performance gains. The consistent improvement across all evaluation metrics—especially recall and AUC-ROC—suggests that the model is well-suited for clinical screening tasks where early identification of positive cases is critical.

The comparative results presented in the baseline and ablation studies reinforce the value of the proposed architecture. Traditional machine learning models performed reasonably well but were limited by their shallow structure and inability to capture non-linear interactions across modalities. Deep learning baselines such as CNN and BiLSTM, while better, still lacked the capability to process multi-source data with sufficient flexibility. The early fusion DNN improved performance, but the multi-input design delivered the most robust and generalizable results. These findings suggest that independent processing of modalities followed by selective integration is more effective than direct feature-level fusion in heterogeneous clinical contexts. The comparison with a multimodal BiLSTM baseline further confirms that the proposed architecture achieves superior performance while maintaining lower model complexity.

A comparison with recent multimodal diagnostic systems highlights the contribution of the proposed framework. Recent multimodal studies have shown that diagnostic performance can be improved by integrating complementary signals across modalities. For example, Cheng et al. proposed a multimodal sleep-disorder detection framework that combines EEG, ECG, and EMG signals for joint sleep-stage and sleep-disorder classification, demonstrating the effectiveness of modality-specific processing in physiological sleep analysis [22]. Similarly, Dentamaro et al. investigated multimodal deep learning for prodromal Parkinson's disease detection using the PPMI database and incorporated explainable AI to support model transparency [24]. More recently, Sar et al. introduced a broader multimodal Parkinson's diagnosis

framework that integrates speech, drawing, neuroimaging, and physiological data to improve diagnostic fidelity [25]. In comparison with these systems, the proposed model focuses on a more accessible combination of non-invasive vocal biomarkers and routinely obtainable sleep-related clinical features. Although the proposed framework uses fewer and lower-cost modalities than imaging- or sensor-intensive systems, it still achieves strong predictive performance while preserving practical feasibility for screening-oriented clinical settings.

From a practical standpoint, the high recall and AUC-ROC values achieved by the proposed model imply its suitability for early screening and decision-support tools. Its ability to handle non-invasive data such as voice recordings and routinely collected physiological measures makes it both scalable and accessible in real-world healthcare settings. Moreover, the architecture can be extended to other multimodal diagnostic tasks beyond neurological disorders by adapting the input branches to different feature domains.

The choice of a multi-layer perceptron (MLP)-based dual-branch architecture is motivated by the nature of the input data used in this study. Both the vocal and clinical sleep datasets are provided in structured tabular form with pre-extracted features, rather than raw temporal signals. As a result, the temporal dynamics of the underlying signals are not available for sequence modeling. In this setting, MLP-based architectures are well-suited for capturing nonlinear relationships within fixed-length feature representations.

Furthermore, the relatively small size of the clinical sleep dataset makes the use of complex temporal models such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks prone to overfitting. The proposed architecture prioritizes robustness and generalization by employing a lightweight yet effective model design. While temporal models may offer additional benefits when raw sequential data are available, the current framework is aligned with the structure and scale of the datasets used in this study.

Limitations

Despite the strong performance of the proposed model, several limitations remain. First, the size of the clinical sleep dataset is relatively small, which may limit generalizability to broader populations. While the vocal dataset is larger, it lacks temporal and longitudinal information that could provide additional diagnostic insights for progressive neurological conditions.

Second, the two datasets used in this study originate from independent cohorts and are not subject-matched. As a result, the model performs feature-level multimodal fusion rather than subject-level integration, which may limit its ability to capture cross-modal relationships at the individual level.

Third, although the proposed architecture is effective, it does not incorporate temporal modeling or attention mechanisms, which could further enhance performance and interpretability.

Finally, while SHAP-based interpretability has been included, additional validation using larger and more diverse clinical datasets is required to confirm the robustness and practical applicability of the model in real-world healthcare settings.

To further enhance the proposed framework, future work will focus on the following directions:

- **Attention-Based Fusion:** Integrate attention mechanisms into the fusion layer to enable the model to weigh modality contributions dynamically, improving interpretability and performance.
- **Temporal Modeling:** Explore temporal deep learning architectures such as recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models to handle time-series data like EEG signals and sequential voice recordings.
- **Transfer Learning and Domain Adaptation:** Apply transfer learning strategies to leverage pretrained models and investigate domain adaptation techniques to improve generalizability across varied clinical populations and data sources.

- **Explainable AI Integration:** Incorporate explainability methods such as SHAP, Local Interpretable Model-agnostic Explanations (LIME), or integrated gradients to interpret model decisions, enhance transparency, and build trust among clinicians.
- **Clinical Validation and Deployment:** Extend the model to real-world settings through collaboration with healthcare institutions, enabling validation on larger, diverse datasets and assessing usability in routine screening workflows.
- **Multimodal Expansion:** Investigate the inclusion of additional modalities such as gait analysis, hand-writing data, or wearable sensor signals to further enhance the diagnostic potential of the framework.

6 Conclusions

This paper proposes a multi-input deep neural network architecture for early diagnosis of neurological disorders by integrating vocal biomarkers and clinical sleep-related features. Unlike traditional models that rely on single-source data or simple feature concatenation, the proposed model processes each modality through dedicated pathways, enabling effective learning of modality-specific patterns and their interactions. Experimental results across multiple metrics demonstrate that this architecture consistently outperforms classical machine learning models and deep learning baselines, with notable gains in recall and AUC-ROC, demonstrating its suitability for early screening applications. The model's ability to use non-invasive and easily collectible data makes it a practical solution for scalable deployment in clinical and remote monitoring contexts. While the results are promising, further validation on larger and more diverse datasets is needed, along with exploring temporal modeling and explainability enhancements. This work contributes a generalizable and extensible framework for multimodal medical diagnosis, offering a pathway toward more accessible and accurate neurological screening tools.

Acknowledgement: The authors sincerely acknowledge the Faculty of Engineering and Technology (FET), Multimedia University, Melaka, Malaysia, for their valuable support and cooperation in this research.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Md. Shujan Shak: Conceptualization, Methodology, Writing—original draft. Nabila Rahman: Data curation, Formal analysis, Visualization. Fuad Mahmud: Software, Validation, Writing—original draft. Ashim Chandra Das: Investigation, Resources, Writing—original draft. M. F. Mridha and Md. Jakir Hossen: Supervision, Writing—review & editing. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available on Kaggle: the Parkinson's Disease Voice Dataset at <https://www.kaggle.com/datasets/ucimachinelearning/parkinsons-voice-dataset> and the Clinical Sleep Dataset at <https://www.kaggle.com/datasets/yfrite/polysom/data>.

Ethics Approval: Ethical approval and consent to participate, consent for publication statement.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Thangaleela S, Sivamaruthi BS, Kesika P, Mariappan S, Rashmi S, Choeisoongnarn T, et al. Neurological insights into sleep disorders in Parkinson's disease. *Brain Sci.* 2023;13(8):1202. doi:10.3390/brainsci13081202.
2. Feigin VL, Nichols E, Alam T, Bannick MS, Beghi E, Blake N, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 2019;18(5):459–80. doi:10.3410/f.735346535.793561284.

3. Steinmetz JD, Seeher KM, Schiess N, Nichols E, Cao B, Servili C, et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* 2024;23(4):344–81. doi:10.2139/ssrn.5200004.
4. Bloem BR, Okun MS, Klein C. Parkinson's disease. *Lancet.* 2021;397(10291):2284–303.
5. Singh KR, Dash S. Early detection of neurological diseases using machine learning and deep learning techniques: a review. In: *Artificial intelligence for neurological disorders*. Cambridge, MA, USA: Academic Press; 2023. p. 1–24.
6. Nunes LC, Pinheiro PR, Pinheiro MCD, Simão Filho M, Nunes REC, Pinheiro PGCD. Automatic detection and diagnosis of neurologic diseases. *IEEE Access.* 2019;7:29924–41. doi:10.1109/access.2019.2899216.
7. Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: a survey. *IEEE Trans Pattern Anal Mach Intell.* 2023;45(10):12113–32. doi:10.1109/tpami.2023.3275156.
8. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina.* 2020;56(9):455. doi:10.3390/medicina56090455.
9. Yan K, Li T, Marques JAL, Gao J, Fong SJ. A review on multimodal machine learning in medical diagnostics. *Math Biosci Eng.* 2023;20(5):8708–26. doi:10.3934/mbe.2023382.
10. Li Y, Daho MEH, Conze PH, Zeghlache R, Le Boité H, Tadayoni R, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Comput Biol Med.* 2024;177(6):108635. doi:10.1016/j.compbimed.2024.108635.
11. Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: a comprehensive review. *Healthcare.* 2022;10(3):541. doi:10.3390/healthcare10030541.
12. Xu X, Li J, Zhu Z, Zhao L, Wang H, Song C, et al. A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. *Bioengineering.* 2024;11(3):219. doi:10.3390/bioengineering11030219.
13. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res.* 2024;26:e59505. doi:10.2196/59505.
14. Gao X, Ma K, Yang H, Wang K, Fu B, Zhu Y, et al. A rapid, non-invasive method for fatigue detection based on voice information. *Front Cell Dev Biol.* 2022;10:994001. doi:10.3389/fcell.2022.994001.
15. Islam MA, Majumder MZH, Hussein MA, Hossain KM, Miah MS. A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets. *Heliyon.* 2024;10(3):e25469. doi:10.1016/j.heliyon.2024.e25469.
16. Rundo JV, Downey R III. Polysomnography. In: *Handbook of clinical neurology*. Vol. 160. Amsterdam, The Netherlands: Elsevier; 2019. p. 381–92.
17. Roebuck A, Monasterio V, Geder E, Osipov M, Behar J, Malhotra A, et al. A review of signals used in sleep analysis. *Physiol Meas.* 2013;35(1):R1. doi:10.1088/0967-3334/35/1/r1.
18. Markun LC, Sampat A. Clinician-focused overview and developments in polysomnography. *Curr Sleep Med Rep.* 2020;6(4):309–21. doi:10.1007/s40675-020-00197-5.
19. ElMoaqet H, Eid M, Glos M, Ryalat M, Penzel T. Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals. *Sensors.* 2020;20(18):5037. doi:10.3390/s20185037.
20. Hireš M, Gazda M, Drotár P, Pah ND, Motin MA, Kumar DK. Convolutional neural network ensemble for Parkinson's disease detection from voice recordings. *Comput Biol Med.* 2022;141(8):105021. doi:10.1016/j.compbimed.2021.105021.
21. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: a scoping review. *npj Digit Med.* 2022;5(1):171. doi:10.1038/s41746-022-00712-8.
22. Cheng YH, Lech M, Wilkinson RH. Simultaneous sleep stage and sleep disorder detection from multimodal sensors using deep learning. *Sensors.* 2023;23(7):3468. doi:10.3390/s23073468.
23. Hemker K, Simidjievski N, Jamnik M. Healnet: multimodal fusion for heterogeneous biomedical data. *Adv Neural Inf Process Syst.* 2024;37:64479–98. doi:10.52202/079017-2057.

24. Dentamaro V, Impedovo D, Musti L, Pirlo G, Taurisano P. Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database. *Sci Rep.* 2024;14(1):20941. doi:10.1038/s41598-024-70165-4.
25. Sar A, Puri PS, Naz H, Aich S, Choudhury T, Gabralla LA. Multi-modal deep learning framework for early detection of Parkinson's disease using neurological and physiological data for high-fidelity diagnosis. *Sci Rep.* 2025;15(1):34835. doi:10.1038/s41598-025-21407-6.