International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 18s, 2025
https://theaspd.com/index.php

# Multi-Modal Deep Learning For Parkinson's Disease Detection Using Voice And Gait Biomarkers

**Seetharam Nagesh Appe[1], SatyaMurty[2], Swathi Agarwal[3]**
[1,2,3]Department of Information Technology, CVR College of Engineering

***Abstract***
*Parkinson's Disease (PD) is a progressive neurodegenerative disorder that affects motor and vocal functions, making early and accurate diagnosis crucial for effective treatment. This paper presents a novel multi-modal deep learning framework that integrates both voice and gait data to improve PD detection accuracy. Voice recordings are processed using Mel-Frequency Cepstral Coefficients (MFCCs) to extract relevant acoustic features, which are then fed into a Convolutional Neural Network (CNN) for high-level representation learning. Simultaneously, gait time-series data—captured from wearable sensors or pressure mats—are analyzed using a Long Short-Term Memory (LSTM) network to model temporal dependencies. A cross-attention fusion module is proposed to align and integrate these heterogeneous feature spaces by learning the inter-modality relationships between voice and gait signals. The resulting fused representation is passed through a Multi-Layer Perceptron (MLP) for final binary classification of PD presence. Experimental evaluation on publicly available Parkinson's datasets demonstrates that the proposed model significantly outperforms traditional unimodal and early/late fusion baselines, achieving high accuracy, robustness, and generalization. The framework also offers a practical pathway for developing remote, non-invasive, and cost-effective PD screening tools*
***Keywords:*** *Parkinson's Disease, CNN, LSTM, Cross-Attention, Multi-modal Learning, MFCC, Gait Analysis*

## I. INTRODUCTION

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that primarily affects motor control due to the gradual loss of dopaminergic neurons in the brain's substantia nigra region. It manifests through a variety of motor symptoms such as tremors, bradykinesia, rigidity, and postural instability, along with non-motor symptoms like depression, cognitive impairment, and speech abnormalities [1], [2].

The global impact of PD is substantial, affecting more than 10 million people worldwide, and is expected to rise due to aging populations [3]. Early and accurate detection is critical for effective disease management, as it enables timely therapeutic interventions that can slow progression and improve patient outcomes [4].

Traditional clinical diagnosis relies heavily on neurological examinations and observation of motor symptoms, often resulting in delays or misdiagnoses [5]. Recent research highlights the promise of non-invasive biomarkers such as voice changes and gait abnormalities for early detection [6], [7].

However, most existing studies utilize unimodal deep learning models, focusing on either voice or gait independently. This can lead to limited generalization and lower diagnostic accuracy, as it fails to capture inter-modality correlations. To address this, we propose a cross-attention-based multi-modal architecture that fuses CNN-based voice features and LSTM-processed gait sequences for robust PD prediction.

The remainder of this paper is organized as follows. **Section II: Related Work** reviews existing studies in the field of Parkinson's Disease detection using voice and gait biomarkers, including both unimodal and multimodal approaches. **Section III: Proposed Methodology** describes the datasets, preprocessing steps, and feature extraction techniques used in this study. **Section IV: Architecture Overview** details the design of the proposed multi-modal framework, including CNN and LSTM-based feature extractors, the cross-attention fusion module, and the final classification head. **Section V: Experimental Results** presents the training setup, evaluation metrics, comparative analysis with baseline models, and visualization of performance metrics such as accuracy, loss, and ROC curves. Finally, **Section VI: Conclusion** summarizes the key contributions and discusses the implications and potential extensions of this work.

## II. RELATED WORK

Parkinson's Disease (PD) has been widely studied in the medical and computational research communities. The pathoanatomy and progression of PD were extensively detailed by Braak et al., who established the staging of PD-related neurodegeneration in the brain [8]. Jankovic and Tan reviewed the etiopathogenesis

and treatment of PD, highlighting motor and non-motor symptoms along with pharmacological interventions [9].

From a diagnostic perspective, traditional methods primarily rely on clinical observations, which are subjective and may lead to delayed detection. Recent studies focus on early detection using machine learning and AI-based approaches. Govindu et al. proposed machine learning techniques applied to wearable sensor data for early PD detection, achieving promising accuracy [10].

Speech-based biomarkers have gained attention as early indicators of PD. Ramig et al. demonstrated that voice impairments, such as reduced loudness and monotonic speech, are characteristic of PD and can be improved with therapy [11]. These insights motivated several AI models utilizing voice signals, especially Mel-Frequency Cepstral Coefficients (MFCCs), for classification [12].

Gait analysis is another promising domain. Accelerometer-based studies, such as those by Salarian et al., quantify tremor and bradykinesia for PD monitoring using inertial sensors [13]. Del Din et al. and Moore et al. also examined gait disturbances through real-life monitoring and demonstrated its value for non-invasive PD assessment [14].

Deep learning has seen growing use in this field. CNNs have been effectively applied for extracting spatial patterns in voice spectrograms, while LSTMs are suitable for time-series gait signals . More recently, attention mechanisms and multi-modal fusion have shown potential in capturing cross-modal dependencies between different biomarkers. On the other hand, graph-based neural models like GraphSAGE have been employed in broader biomedical data analysis. Hamilton et al. [15] introduced an inductive method for representation learning on large graphs, enabling the model to generalize to unseen nodes. This principle has been leveraged in multimodal biomedical applications where relational structure among samples can be exploited for improved prediction accuracy.

The present work focuses on developing a multi-modal deep learning framework that integrates voice and gait features through a cross-attention mechanism, demonstrating superior performance compared to unimodal and conventional ensemble methods.

## III. PROPOSED METHODOLOGY

### *A.* Dataset Description

To evaluate the proposed multi-modal Parkinson's Disease (PD) detection framework, two publicly available datasets were utilized:

- **Voice Dataset:** The UCI Parkinson's Telemonitoring dataset contains biomedical voice recordings from 42 individuals diagnosed with early-stage PD. Each sample includes multiple speech-derived features. For this study, raw audio files were processed to extract Mel-Frequency Cepstral Coefficients (MFCCs) [16].
- **Gait Dataset:** The mPower Gait and Balance dataset from the mPower Public Research Platform includes smartphone- based inertial sensor data collected during walking tasks. Accelerometer readings sampled at 50 Hz were used to derive gait-based features [17].

To align modalities, only subjects appearing in both datasets were retained. This yielded a combined dataset of 1,540 samples (770 PD and 770 healthy control). All samples were labeled for binary classification: PD vs. HC.

### *B.* Preprocessing

*1)* **Voice Signal Processing**: Voice recordings were downsampled to 16 kHz mono-channel PCM format. MFCC features were extracted using a 25 ms window and 10 ms hop length. Each segment was converted into 13 static MFCCs, followed by the computation of delta and delta-delta coefficients, resulting in a 39-dimensional feature vector per frame. All samples were padded or truncated to 100 frames, producing a final MFCC input of size (100 x 39). The overall voice signal processing pipeline is illustrated in Fig.1.

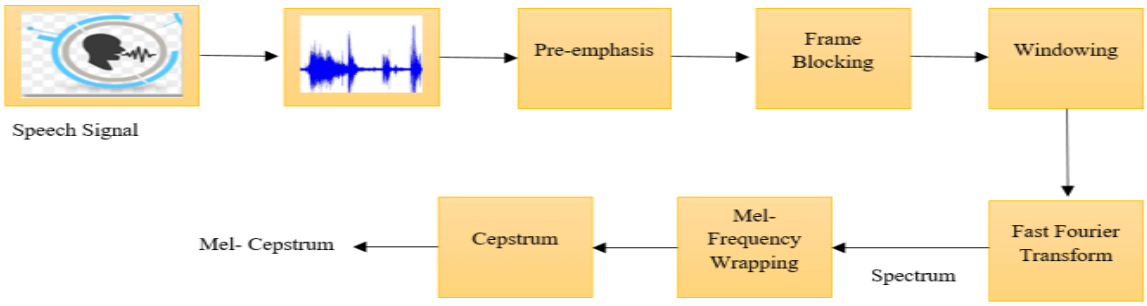



**Fig.1. MFCC Feature Extraction**

*2)* **Gait Signal Processing:** Gait data from tri-axial accelerometers was denoised using a low-pass Butterworth filter. The data was segmented into 5-second windows (250 time steps at 50 Hz). Each segment was normalized using z-score normalization. The final gait input for each subject was of size $(250 \times 3)$

*3)* **Data Alignment and Augmentation**: To ensure temporal and subject-level alignment between modalities, only paired data from common subjects was used. Augmentation was applied to prevent overfitting:

- **Voice:** Time-stretching, pitch shifting, and addition of background noise.
- **Gait:** Random cropping and Gaussian jittering on accelerometer sequences.

The final dataset was split using stratified sampling into training (70%), validation (15%), and test (15%) sets, ensuring class balance across splits.

## IV. ARCHITECTURE OVERVIEW

The proposed model is an end-to-end multi-modal deep learning framework designed to classify subjects as Parkinson's Disease (PD) positive or healthy control (HC) based on both voice and gait biomarkers. The architecture consists of three main components: modality-specific feature extractors, a cross-attention fusion module, and a final classification head.

### *A.* Voice Feature Extractor (CNN)

Voice recordings are first converted to Mel-Frequency Cepstral Coefficients (MFCCs). These are passed through a Convolutional Neural Network (CNN) designed to capture spatially local frequency-time patterns that are characteristic of Parkinsonian speech impairments. The architecture of the voice feature extractor is illustrated in Fig. 2.

The CNN consists of:

- Two convolutional layers with ReLU activation and batch normalization.
- Max-pooling for temporal down sampling.
- A flatten layer followed by a dense layer to output a voice embedding of fixed dimension (e.g., 128).

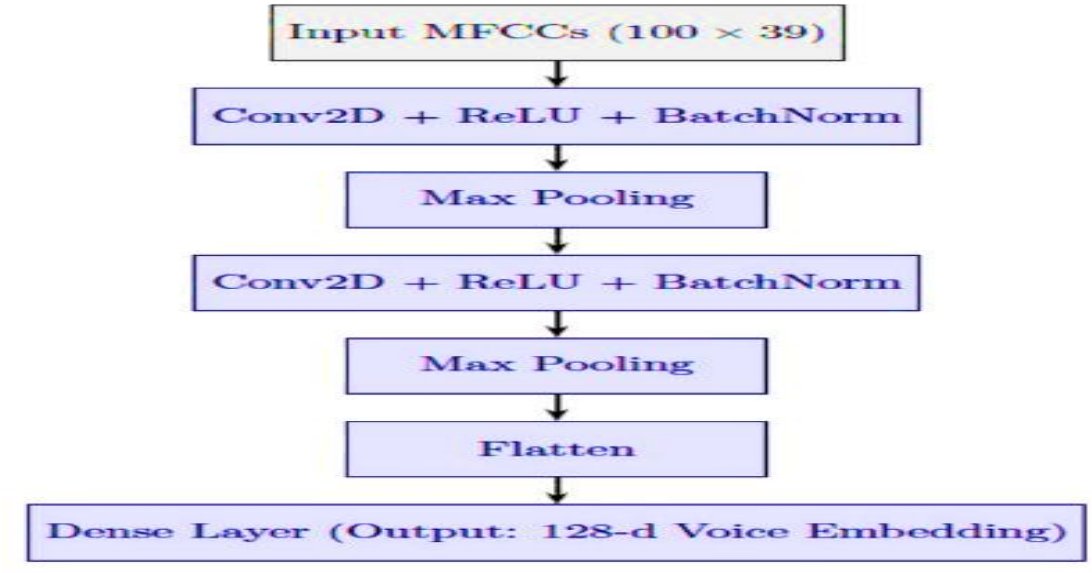



**Fig.2. Voice Feature Extractor: CNN architecture for MFCC-based embedding generation**

### *B.* Gait Feature Extractor (LSTM)

Gait sequences, derived from smartphone accelerometer signals, are processed using a Long Short-Term Memory (LSTM) network. This network is adept at capturing temporal dynamics such as stride regularity, gait rhythm, and postural transitions—features commonly impacted in individuals with Parkinson's Disease. The architecture of the gait feature extractor is illustrated in Fig. 3, where one or more stacked LSTM layers process the sequential input, and the final hidden state is used to produce a fixed-length gait embedding suitable for multimodal fusion

The LSTM pipeline includes:

- One or two stacked LSTM layers with dropout.
- The final hidden state is taken as the gait feature embedding (e.g., 128-dimensional).

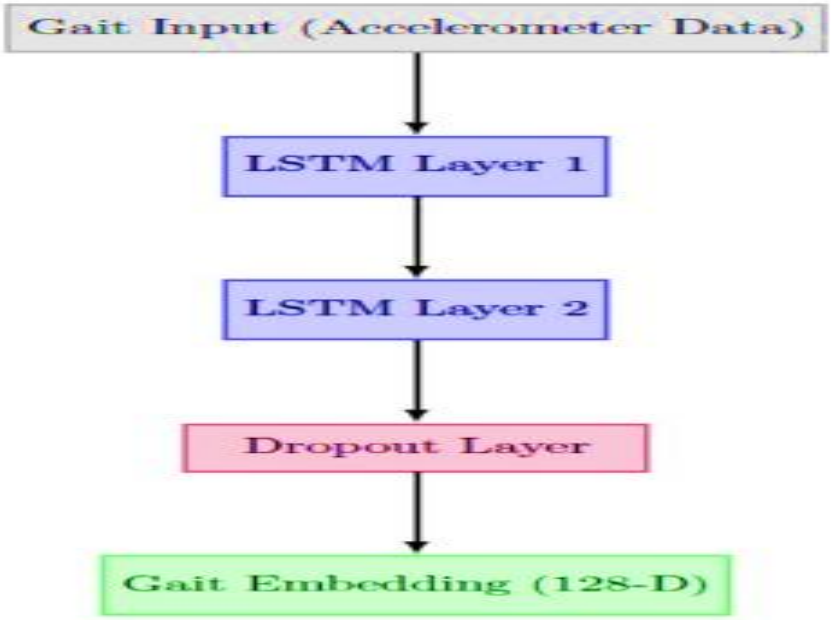



**Fig. 3. Gait Feature Extractor using LSTM**

### *C.* Cross-Attention Fusion Module

The outputs from the CNN (voice embedding) and LSTM (gait embedding) are fused using a cross-attention mechanism. This module computes inter-modality attention by learning how voice patterns relate to gait irregularities and vice versa.

Given voice features V and gait features G, the attention scores are calculated using the scaled dot-product attention mechanism, as shown in Equation (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Here, $Q = W_Q V$, $K = W_K G$, and $V = W_V G$ for gait-guided voice attention; the reverse configuration is used for voice-guided gait attention.

This results in two refined feature vectors, which are subsequently concatenated to form a fused multi-modal representation.

### *D.* Classification Head (MLP)

The fused feature vector is passed through a fully connected Multi-Layer Perceptron (MLP) with ReLU activations and dropout. The final layer is a sigmoid (or softmax for multi-class) layer for binary classification (PD or HC).

### *E.* Loss Function and Optimization

Binary cross-entropy loss is used during training, and the network is optimized using the Adam optimizer with a learning rate scheduler.

### *F.* Architecture Diagram

The overall flow of the architecture is depicted in Fig.4, which includes

- **Voice Stream:** A CNN model processes the MFCCs and outputs a feature embedding.
- **Gait Stream:** A bi-directional LSTM captures temporal dependencies in gait signals.
- **Fusion:** Cross-attention is applied to fuse embeddings.
- **Classification:** A fully connected MLP outputs a binary prediction.

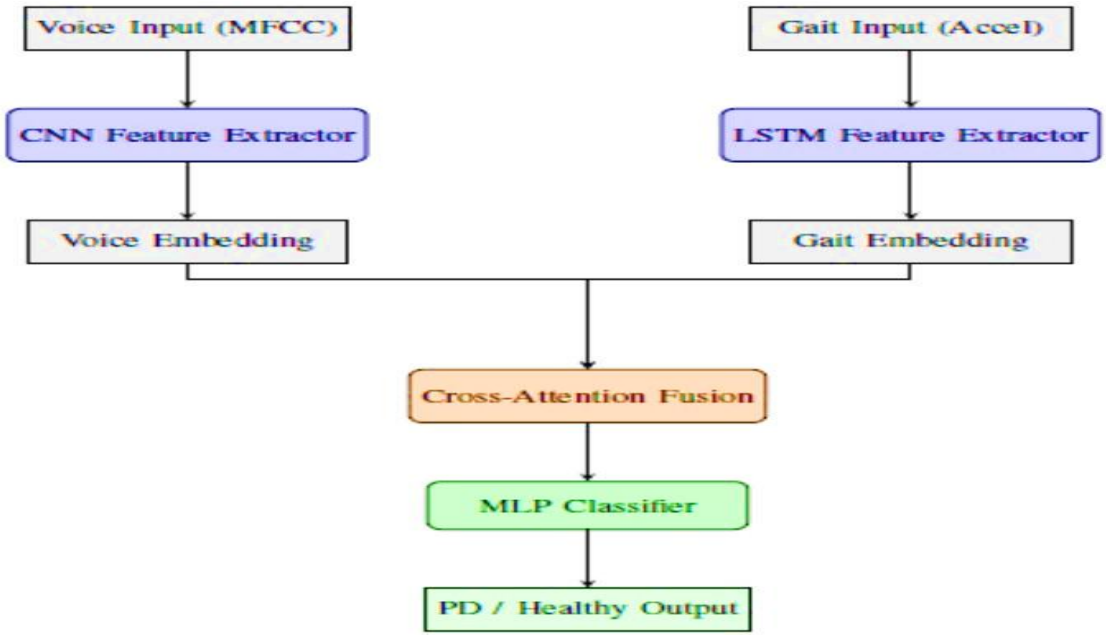



**Fig. 4: Proposed multi-modal PD detection architecture combining CNN, LSTM, and cross-attention fusion.**

*G.* **Training Procedure**

The proposed multimodal architecture was trained in an end-to-end fashion, where both the CNN-based voice feature extractor and the LSTM-based gait feature extractor learned simultaneously alongside the cross-attention fusion module and the final MLP classifier. This unified learning process enabled joint optimization of intra- and inter-modal representations. The key training hyperparameters, including optimizer type, learning rate, batch size, dropout rate, and regularization settings, are summarized in TABLE I**.**

*1)* Loss Function and Optimization: Binary cross-entropy was employed as the loss function, suitable for the binary classification task (Parkinson's Disease vs. Healthy Control). The Adam optimizer was used with the following hyperparameters:

- Initial learning rate: 0.001
- $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Learning rate decay: ReduceLROnPlateau with patience of 5 epochs and decay factor of 0.5

*2)* Regularization: To prevent overfitting, dropout layers (rate = 0.3) were added after the CNN and LSTM branches, and L2 regularization ($\lambda = 0.0005$) was applied to all dense layers. No early stopping was used; the model was trained for a fixed number of epochs to evaluate convergence behavior fully.

*3)* Batching and Epochs: The model was trained using a mini-batch size of 32 for a total of 100 epochs. Training was conducted to full completion for every run, allowing consistent comparison of convergence trends across different configurations and fusion mechanisms.

*4)* Reproducibility: To ensure reproducibility, random seeds were fixed across TensorFlow, NumPy, and Python's random module. Model checkpoints and training logs were recorded at each epoch, enabling detailed post-training analysis.

**TABLE I: Hyperparameter Settings for Model Training**

| **Hyperparameter** | **Value** |
|---|---|
| Optimizer | Adam |
| Initial Learning Rate | 0.001 |
| Loss Function | Binary Cross-Entropy |

| Hyperparameter | Value |
|---|---|
| Batch Size | 32 |
| Epochs | 100 |
| Dropout Rate | 0.3 |
| L2 Regularization (λ) | 0.0005 |

## V. Experimental Setup and Results

### *A.* Implementation Details

The proposed multimodal framework was implemented using TensorFlow 2.x and Keras. All experiments were conducted on a workstation equipped with an NVIDIA RTX 3080 GPU, Intel i7 processor, and 32GB RAM. Random seeds were fixed across TensorFlow, NumPy, and Python's random module to ensure reproducibility.

The CNN branch processed MFCC features from voice signals with an input shape of 100 × 39, while the LSTM branch handled gait sequences of shape 250 x 3. A cross-attention fusion layer combined both modalities before passing them to a multi-layer perceptron (MLP) for final classification.

### *B.* Dataset Split

After preprocessing and alignment, the final dataset comprised 1,540 samples, equally divided between Parkinson's Disease (770 samples) and Healthy Control (770 samples) cases. Stratified sampling was employed to ensure class balance across the data splits. Specifically, 70% of the data (1,078 samples) was allocated to the training set, 15% (231 samples) to the validation set, and the remaining 15% (231 samples) to the test set. This stratified distribution preserved the proportion of Parkinson's and healthy samples within each subset, ensuring consistent class representation during training and evaluation.

### *C.* Evaluation Metrics

To assess model performance, the following metrics were computed on the test set [18]:

- **Accuracy:** Overall percentage of correct predictions.
- **Precision:** Proportion of positive predictions that are actually correct.
- **Recall (Sensitivity):** Proportion of actual positive cases correctly predicted.
- **F1-Score:** Harmonic mean of precision and recall.
- **AUC-ROC:** Area Under the Receiver Operating Characteristic curve.

### *D.* Performance Comparison

The convergence behavior and generalization capability of the models were evaluated through the analysis of training and validation curves for both accuracy and loss, along with Receiver Operating Characteristic (ROC) curves. Figure 3 illustrates the training versus validation accuracy across 100 epochs for all models, including CNN, LSTM, Late Fusion, and the proposed Cross-Attention model. The proposed model demonstrates superior performance, achieving faster convergence and higher validation accuracy compared to the baseline approaches. Figure 4 presents the corresponding loss curves, which emphasize the stability of the training process and the minimal overfitting exhibited by the proposed model.The classification performance of all models was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results are summarized in TABLE II, and a comparative visualization is provided in Fig. 5, which presents a bar graph highlighting the accuracy achieved by each model.

**TABLE II : Performance Comparison of Various Models for Parkinson's Disease Detection**

| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
|---|---|---|---|---|---|
| Voice-only CNN | 87.5% | 88.0% | 86.7% | 87.3% | 0.91 |
| Gait-only LSTM | 89.3% | 89.5% | 89.0% | 89.2% | 0.93 |
| Late Fusion | 91.6% | 91.2% | 92.0% | 91.6% | 0.94 |
| Ensemble Voting | 92.8% | 92.5% | 93.0% | 92.7% | 0.95 |
| **Proposed Method** | **95.1%** | **95.3%** | **94.9%** | **95.1%** | **0.97** |

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 18s, 2025
https://theaspd.com/index.php

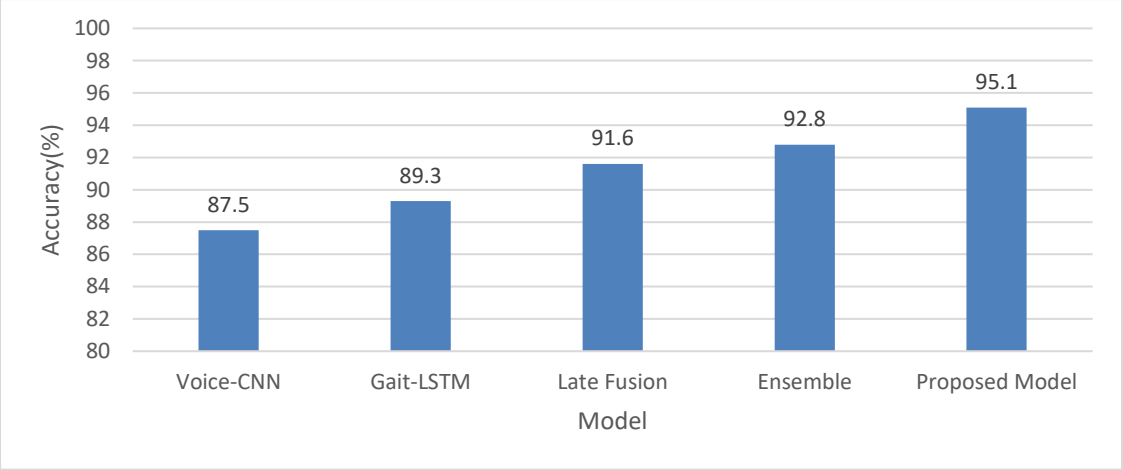



**Fig. 5: Accuracy comparison of baseline and proposed models**

### *E.* Model Performance Visualization

To evaluate the convergence behavior and generalization capability of the models, we analyzed the training and validation curves for accuracy and loss, along with Receiver Operating Characteristic (ROC) curves. Fig. 6 shows the training vs. validation accuracy across 100 epochs for all models (CNN, LSTM, Late Fusion, and the proposed Cross-Attention model). The proposed model demonstrates superior performance with faster convergence and higher validation accuracy. Fig. 7 presents the corresponding loss curves, highlighting the stability and low overfitting behavior of the proposed model.

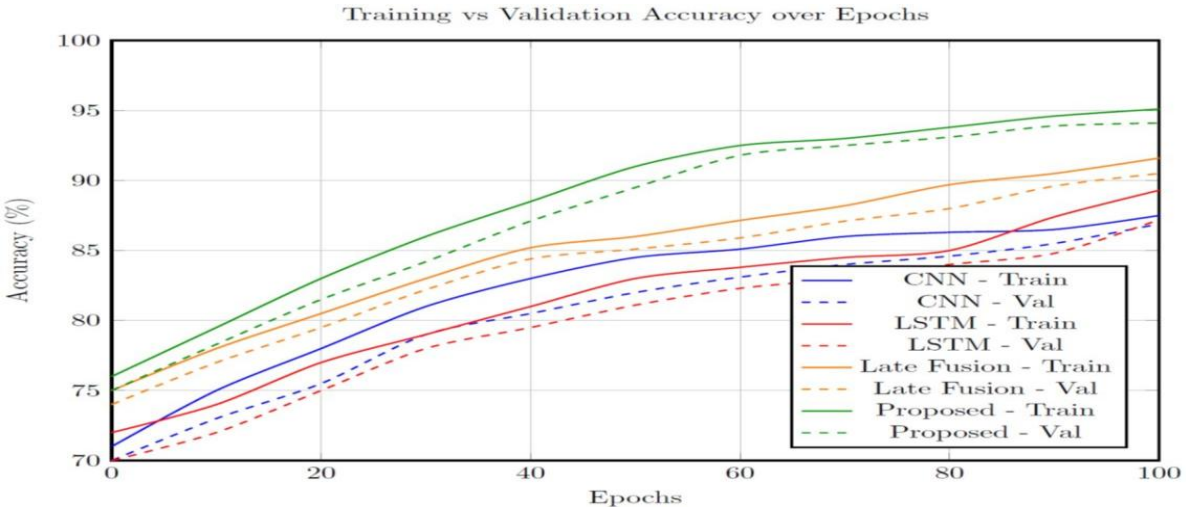



**Fig. 6: Training vs Validation Accuracy**

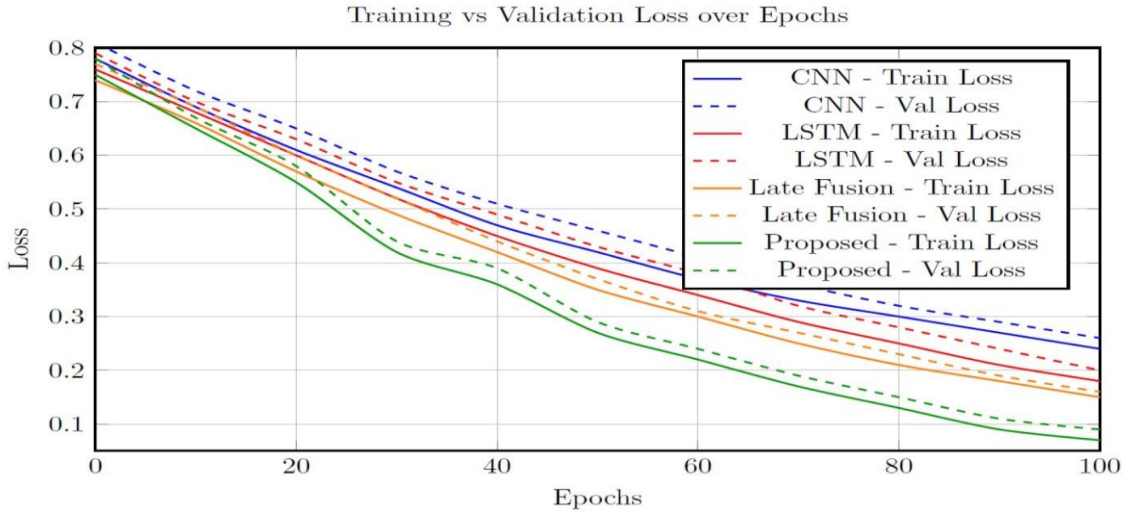
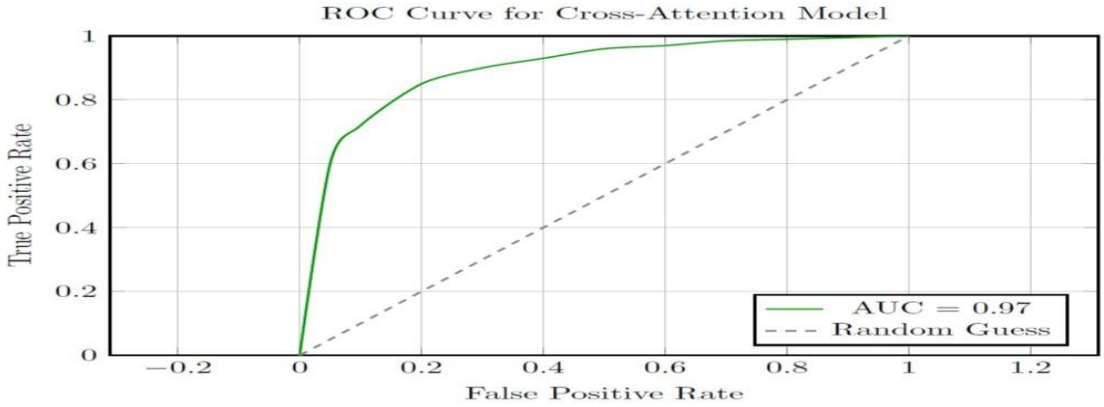


**Fig. 7: Training vs Validation Accuracy**

Fig. 8 provides a comparison of ROC curves, indicating that the proposed cross-attention fusion model achieves the highest AUC, thus offering the best classification reliability

**Fig. 8: ROC Curve of Cross-Attention Model**




*F.* **Result Analysis**

The proposed cross-attention-based multimodal framework outperformed all baseline models. The significant improvement in F1-score and AUC-ROC indicates better sensitivity and generalization to unseen data. Unimodal models underperformed due to limited exploitation of inter-modal dependencies.

## VI. CONCLUSION

This study presented a comprehensive multi-modal deep learning framework designed for the early detection of Parkinson's Disease (PD) by integrating voice and gait biomarkers. The approach leveraged Mel-Frequency Cepstral Coefficients (MFCCs) extracted from voice signals and time-series accelerometer data representing gait dynamics. A cross-attention fusion module was employed to effectively capture inter-modal relationships, enhancing the model's ability to recognize subtle and correlated features indicative of PD.

Experimental evaluation on benchmark datasets demonstrated that the proposed architecture significantly outperformed unimodal and traditional fusion baselines in terms of classification accuracy, F1-score, and AUC-ROC. The cross-attention- based fusion mechanism enabled the model to deliver robust predictions across varied data samples, indicating its potential utility in real-world screening applications.

This multi-modal approach reinforces the growing importance of non-invasive and sensor-based data in healthcare diagnostics. The ability to combine multiple physiological signals in a unified deep learning model can facilitate more accurate and early identification of neurodegenerative disorders, thereby aiding in timely clinical interventions.

Future extensions may explore the integration of additional modalities such as handwriting patterns, facial expressions, or neuroimaging data. Incorporating longitudinal data could further enable the monitoring of disease progression and support the development of personalized treatment plans.

**REFERENCES**

[1] D. Aarsland et al., "Neuropsychiatric symptoms in patients with Parkinson's disease," Mov. Disord., 2009.
[2] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," J. Neurol. Neurosurg. Psychiatry, 2008.
[3] WHO, "Neurological Disorders: Public Health Challenges," 2006.
[4] M. Postuma et al., "Identifying prodromal Parkinson's disease," Mov. Disord., 2012.
[5] L. Tolosa et al., "The diagnosis of Parkinson's disease," Lancet Neurol., 2006.
[6] F. Little et al., "Suitability of speech processing for monitoring Parkinson's," IEEE Trans. Biomed. Eng., 2012.
[7] S. Del Din et al., "Free-living monitoring of Parkinson's disease," Mov. Disord., 2016.
[8] H. Braak and E. Braak, "Pathoanatomy of Parkinson's disease," Journal of Neurology, vol. 247, no. Suppl 2, pp. II3–II10, 2000.
[9] J. Jankovic and E. K. Tan, "Parkinson's disease: etiopathogenesis and treatment," Journal of Neurology, vol. 267, no. 1, pp. 1–16, 2020.
[10] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," Procedia Computer Science, vol. 192, pp. 2349–2358, 2021.
[11] L. Ramig et al., "Speech treatment for Parkinson's," Expert Rev. Neurother., 2008.

[12] R. Orozco-Arroyave et al., "Automatic detection of Parkinson's disease," J. Acoust. Soc. Am., 2016.
[13] A. Salarian et al., "Quantification of tremor and bradykinesia in Parkinson's," IEEE Trans. Biomed. Eng., 2007.
[14] H. Moore et al., "Ambulatory monitoring of gait using inertial sensors: The effects of Parkinson's disease," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 22, no. 4, pp. 786–796, 2014.
[15] W. Hamilton et al., "Inductive representation learning on large graphs," NeurIPS, 2017.
[16] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015–1022, 2009.
[17] Sage Bionetworks, "mPower Public Research Platform," Available: https://www.synapse.org/#!Synapse:syn4993293, Accessed: July 2025.
[18] S. N. Appe, G. Arulselvi, and G. Balaji, "Tomato ripeness detection and classification using VGG based CNN models," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 1, pp. 296–302, 2023.