information

MDPI

Article

PD-Net: Parkinson's Disease Detection Through Fusion of Two Spectral Features Using Attention-Based Hybrid Deep Neural Network

Munira Islam [1], Khadija Akter [1], Md. Azad Hossain [1,*] and M. Ali Akber Dewan [2,*]

[1] Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering & Technology, Chattogram 4349, Bangladesh; islammunira56@gmail.com (M.I.); khadija.ete@cuet.ac.bd (K.A.)
[2] School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Athabasca, AB T9S 3A3, Canada
* Correspondence: azad@cuet.ac.bd (M.A.H.); adewan@athabascau.ca (M.A.A.D.)

Abstract: Parkinson's disease (PD) is a progressive degenerative brain disease that worsens with age, causing areas of the brain to weaken. Vocal dysfunction often emerges as one of the earliest and most prominent indicators of Parkinson's disease, with a significant number of patients exhibiting vocal impairments during the initial stages of the illness. In view of this, to facilitate the diagnosis of Parkinson's disease through the analysis of these vocal characteristics, this study focuses on exerting a combination of mel spectrogram and MFCC as spectral features. This study adopts Italian raw audio data to establish an efficient detection framework specifically designed to classify the vocal data into two distinct categories: healthy individuals and patients diagnosed with Parkinson's disease. To this end, the study proposes a hybrid model that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) for the detection of Parkinson's disease. Certainly, CNNs are employed to extract spatial features from the extracted spectro-temporal characteristics of vocal data, while LSTMs capture temporal dependencies, accelerating a comprehensive analysis of the development of vocal patterns over time. Additionally, the merging of a multi-head attention mechanism significantly enhances the model's ability to concentrate on essential details, hence improving its overall performance. This unified method aims to enhance the detection of subtle vocal changes associated with Parkinson's, enhancing overall diagnostic accuracy. The findings declare that this model achieves a noteworthy accuracy of 99.00% for the Parkinson's disease detection process.

Keywords: Parkinson disease; degenerative; spectral features; hybrid model; multi-head attention

Academic Editor: Heming Jia

Received: 13 January 2025
Revised: 29 January 2025
Accepted: 5 February 2025
Published: 12 February 2025

Citation: Islam, M.; Akter, K.; Hossain, M.A.; Dewan, M.A.A. PD-Net: Parkinson's Disease Detection Through Fusion of Two Spectral Features Using Attention-Based Hybrid Deep Neural Network. *Information* **2025**, *16*, 135. https://doi.org/10.3390/info16020135

1. Introduction

The well-known and intricate neurological disorder known as Parkinson's disease is characterized by malfunctions in motor and non-motor systems affecting people 60 years or older and is generated due to the loss of dopaminergic neurons along with Lewy bodies in the midbrain [1]. PD is a complicated medical condition to treat as its wide range of symptoms including postural instability, tremors, rigidity, dysarthria, hypomimia, dysphagia, and cognitive impairment. Studies highlight that, by 2040, there will be 12.9 million instances of Parkinson's disease, which is predicted to quadruple from over 6.2 million in 2015, making effectual identification even more necessary [2]. However, still, there is no specific diagnostic procedure for this; diagnosing the condition is problematic and depends

on clinical symptoms and criteria. The main goal of a PD detection method is to deploy diverse techniques to assess symptom severity. Vocal dysfunction is one of the earliest and most prominent signs so many patients show vocal defects early in the illness [3]. Often, speech abnormalities, appearing before any visible movement issues, provide vital early indicators for PD diagnosis. The need for comprehensive screening, including speech recognition technology, is highlighted by non-vocal challenges like memory loss and anxiety.

Before machine learning, PD diagnosis depended totally on clinical assessments by specialists using tools like the Unified Parkinson's Disease Rating Scale (UPDRS) [4]. Nevertheless, the inception of machine learning has revolutionized Parkinson diagnosis. Speech analysis, specifically with machine learning, provides an affordable, non-invasive substitute for expensive imaging techniques. Speech abnormalities, such as changes in rhythm, monotone speech, articulation, and pitch, often emerge early and offer opportunities for early detection [5].

Neural networks distinguish speech patterns as probable biomarkers, providing a non-invasive, economical means of focused therapy and monitoring [6,7]. This work improves the Parkinson's early diagnosis process and adapted treatment by utilizing deep learning models including attention layers while transforming audio recording data into speech-associated feature sets. The principal contribution of this research encompasses the following:

- This study employs the fusion of mel spectrogram and MFCC (mel-frequency cepstral coefficients) as essential spectral features to capture the intricate vocal characteristics associated with Parkinson's disease. Through these features, the effective representation of voice signals is carried out in a way that highlights the differences between healthy individuals and those with Parkinson disease.
- This study establishes a hybrid model that incorporates Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This combination leverages the strengths of CNNs in spatial feature extraction and LSTMs in capturing temporal patterns. By incorporating a multi-head attention mechanism, the model enhances its ability to focus on the most critical features in the vocal data. This improves the overall performance and accuracy by allowing the model to prioritize the most relevant spectral and temporal details for Parkinson's disease detection.
- A comparative analysis of the proposed model against existing works is conducted to assess its performance and effectiveness.

In Section 2, the relevant literature is synthesized, while, in Section 3, background data on deep neural networks and disease classification are provided. In Section 4, the experimental analysis and the results are discussed; Sections 5 and 6 contain the discussion and conclusion, respectively.

2. Related Work

Parkinson's disease is a crucial condition that causes abnormalities in neurons and protein aggregation, which interferes with day-to-day functioning. It is still challenging to achieve a detection accuracy of more than 90% in real-time situations. To address this, various clinical and technical solutions including neuroprotective therapies and molecular interventions have been developed [8,9]. Manual methods including analyzing handwriting, speech, and medical symptoms often lack accuracy in detection. To address this challenge, advanced computational methods like various machine learning and deep learning techniques have been explored [10], with the primary objectives of identifying and predicting Parkinson symptoms. This section provides a brief overview of earlier research on speech analysis in PD identification.

Speech-Based Parkinson Disease Detection

Prior to the development of machine learning, neurologists' clinical judgments were the main method used to diagnose PD. They used instruments such as the Unified Parkinson's Disease Rating Scale (UPDRS) to measure the severity and course of the disease as they noticed motor signs such as tremors, rigidity, bradykinesia (slowness of movement), and postural instability. Rather than being used to directly diagnose PD, imaging techniques like magnetic resonance imaging (MRI) and computed tomography (CT) scans were used to distinguish between other ailments with similar symptoms, like tumors or strokes. However, as the mild first symptoms could vary widely among the patients, these methods were not very reliable, especially at the very first phases of the disease. This frequently produced delaying until the illness reached more noticeable stages, making immediate intervention more difficult.

In recent years, the use of machine learning to classify Parkinson disease has changed magnificently. Notably, by inspecting speech patterns, machine learning has showed major potential in identifying early criteria of Parkinson by revealing minute differences suggestive of the illness [11]. Extensive voice data are easily processed by ML systems, which makes it suitable for them to detect slight variations in vocal characteristics like tone, pitch, and rhythm that may go unobserved in conventional clinical examinations [12]. Other diagnostic techniques, like MRI imaging, offer insights into brain anatomy and physiology but they often fail to detect the early, subtle changes associated with PD. This is because of the fact that MRI basically detects more severe neurodegenerative alterations that emerge in later stages of the disease [13]. Another well-known clinical test is handwritten spirals, which are destined to measure fine motor abilities, which decline in Parkinson's patients. Moreover, a number of variables, like the patient's mental state or outside circumstances, may have consequences on the results, possibly producing subjective evaluations. Though valuable, this method is less accurate and reliable compared to speech analysis [14]. Thus, speech analysis is unique due to the fact that it offers an objective, early, non-intrusive diagnostic tool—a critical component of urgent intervention and treatment.

The nearest neighbor boosting technique was employed on sound signals in .wav format by Aditya et al. [15]. Their approach did not use well-developed deep learning techniques, even though they were able to obtain an accuracy of 0.75. This innovative method, known as nearest neighbor boosting (NNB), was tested on the Parkinson's Speech Dataset, which includes 27 input speech features.

By utilizing machine learning approaches with speech samples at different frequencies based on speech biomarkers, Amran et al. [16] gained progress in the contrary direction. Supervised machine learning algorithms, validated through 10-fold cross-validation, were applied to these data. A notable breakthrough in the field of early Parkinson's disorder identification was made by their method, which produced an accuracy of 0.84.

Aditi et al. [17] examined and contrasted many machine learning approaches for the processing of MVDP audio data. These recordings integrate various speech parameters such as jitter, shimmer, and harmonics-to-noise ratio. Their research yielded an accuracy of 0.91, demonstrating the possibility of adding more voice and REM sleep information for even greater advancement.

Parallel to this, Umesh et al. [18] concentrated on voice signals and CNN-LSTM architecture with efficient hyperparameter tuning for PD detection. Their method achieved a slightly better accuracy of 0.93, outperforming earlier approaches and proving the usefulness of deep learning methods for the diagnosis of Parkinson's disease.

Raya et al. [6] detected Parkinson's based on voice signal features while analyzing the use of machine learning algorithms. Their study revealed the potential of various ML models in achieving a highest classification accuracy of 98.49% with the MLP model.

Gayarathi et al. [19] explored deep learning models, including ADRNN and ADCNN, for the detection of Parkinson's disease from vocal data. Their ADCNN model achieved the best accuracy of 98.32%, exhibiting the competence of deep learning in studying speech features for early diagnosis.

3. Materials and Methods

This section provides a concise description of the dataset utilized in this study, along with a detailed explanation of the working principles of the proposed network. The detailed overview of the proposed Parkinson's disease detection framework is presented in Figure 1.

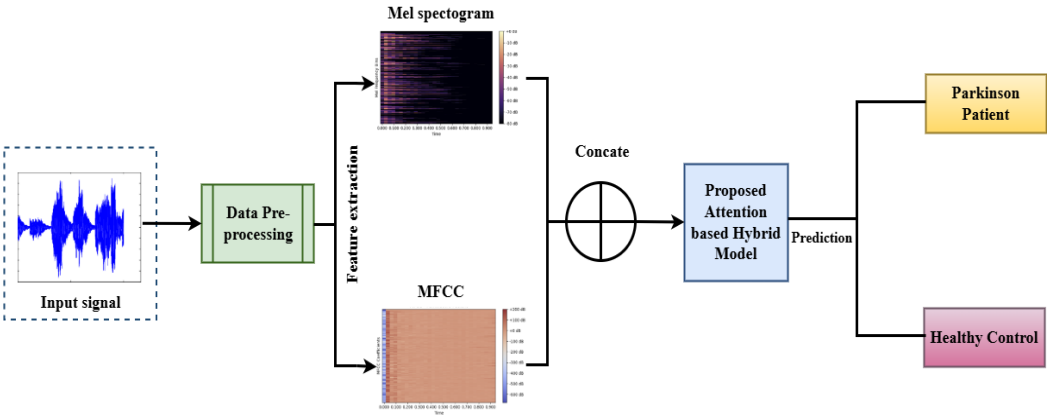


Figure 1. Abstract view of the proposed Parkinson disease detection framework through vocal analysis.

The framework starts by receiving raw audio input signals, which undergo data pre-processing to clean and normalize the vocal data, preparing them for analysis. Given the dataset's initial modest size of 831 samples, augmentation techniques are employed to expand it to 3496 samples, ensuring a balanced distribution across categories to reduce overfitting and enhance generalization.

In the feature extraction stage, two spectral features, mel spectrogram and MFCC (mel-frequency cepstral coefficients), are computed to capture the key characteristics of the voice signal. These features provide a rich representation of both the frequency content (through mel spectrogram) and the short-term power spectrum (via MFCC) of the vocal data, which are crucial for detecting subtle vocal changes associated with Parkinson's disease.

Next, the extracted features are concatenated and fed into the proposed attention-based hybrid model, which consists of a CNN to extract spatial features and a LSTM network to capture temporal dependencies in the vocal patterns. Additionally, L2 regularization is applied to penalize overly complex weights, further reducing the risk of overfitting. A multi-head attention mechanism is incorporated to enhance the model's focus on the most relevant features from the vocal data. The final prediction stage classifies the vocal data into two categories: Parkinson's patient or healthy control, thereby facilitating early and accurate detection of the disease.

3.1. Dataset Description

The raw audio data [20] were used in this study to develop Parkinson disease detection. In this study, words with <4 characters were not analyzed. The very first participants were young, healthy people (YHC) between the ages of 19 and 29, chosen to create a trustworthy baseline for neuromotor traits. Participants stood 15–25 cm away from the microphone

and read aloud in an echo-free environment. A warm setting (22 °C) preceded the start of sessions at 10 a.m. The reading materials were presented in Times New Roman, bold, and with a font size of 20. Included were the following voice recordings:

- Two phonemically balanced text reads separated by a 30 s break;
- Uttering "pa" for 5 s, pausing for 20 s, then "ta" for 5 s;
- Two vocal phonations of "a";
- Two vocal phonations of "e"
- Two vocal phonations of "i";
- Two vocal phonations of "o";
- Two vocal phonations of "u";
- Reading a few phonemically balanced words, pausing for one minute, then reading phonemically balanced phrases.

The distribution of audio data is shown in Figure 2, a pie chart in the dataset used to classify Parkinson's disease. It includes 336 recordings from healthy controls and 495 recordings from Parkinson's sufferers, for a total of 831 audio samples. A fair analysis is ensured by the study design, which takes into consideration the slight imbalance between the patient and healthy control samples. Here, Table 1 provides the number of audio samples in each dataset category that includes original, augmented, and balanced data.

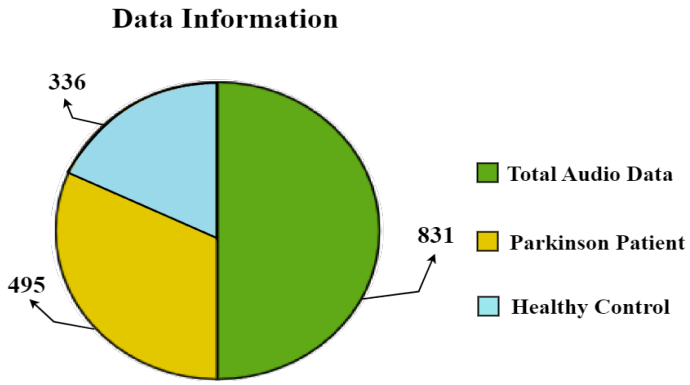


Figure 2. Distribution of audio data in the dataset for Parkinson's disease classification.

Table 1. Number of audio data in each dataset.

Dataset	Total Instances	Parkinson Patient	Healthy Control (HC)
Original audio	831	495	336
Augmented audio	3324	1748	1576
Balanced audio	3496	1748	1748

3.2. Data Preparation

The data preparation part begins with segmenting continuous audio signals into 30 ms frames. As it allows for the analysis of short-time frequency information within the signals, this step is crucial for feature extraction process.

In subsequent segmentation, the audio signals—sampled at 16,000 Hz with a duration of 3 s—are subjected to various augmentation techniques in order to enhance the original dataset. These augmentation methods are precisely selected as they imitate real-world conditions, enhancing models' ability to recognize patterns across diverse audio contexts and improving performance on unseen data.

Random oversampling is applied after augmentation to address class imbalance and ensure fair representation of the data. Then, the upcoming step involves extracting key

auditory features from the audio recordings which provides informative representations of the audio that serve as input for the machine learning model.

3.2.1. Data Augmentation

To develop the resilience and generalization of deep neural networks trained on speech data, augmentation techniques including pitch scaling, random gain modification, and white noise addition are vital [21].

- **Pitch Scaling**: In our study, pitch scaling was used with a factor of 1.5 in the audio dataset to raise the frequency of the audio without changing its duration, therefore maintaining the vocal clarity. By highlighting small changes in vocal patterns associated with PD, this additional variability boosted the model's capacity to learn diverse audio signals linked to the condition. The pitch scaling formula [22] is given below:

$$Pitch\,factor = 1 + \frac{change\ \ in\ \ pitch}{12} \tag{1}$$

  where the degree of pitch alteration is determined by the pitch factor.
- **Random Gain**: The Parkinson's audio data was subjected to random gain modification, which incorporates fluctuating the gain or level within a preset range of 10%. By regulating the gain factor between 0.9 and 1.1, the loudness was carefully fluctuated. This technique of augmentation improves the model's capacity to generalize by imitating various volume levels that correspond to actual acoustic situations.
- **Addition of White Noise**: White noise, which is random and has a flat frequency spectrum, was added to the Parkinson's audio data with a noise factor of 0.1. This increases the model's resistance to background interference by simulating ambient noise. To guarantee legitimate noise intensity, the noise factor must stay non-negative.

3.2.2. Data Balancing

To achieve a balanced representation between Parkinson's patients and healthy controls in the dataset, the random oversampling technique is used. This method generates additional synthetic samples for the minority class (Parkinson's patients) by replicating existing examples until the number of samples matches the majority class (healthy controls) or meets a defined sampling ratio [23]. By addressing class imbalance, this approach helps the model learn effectively from both groups, resulting in improved class distribution and better performance in machine learning tasks.

3.3. Feature Extraction

- **Extraction of Spectrograms**: Speech recordings throughout time are subjected to a spectrogram analysis, shown in Figure 3, in order to categorize Parkinson's cases. Audio is split into shorter frames for frequency extraction using the Fourier transform after a window is applied to minimize spectral leakage. This creates a time–frequency matrix, where rows stand for frequency bins and columns for time frames. Subtle speech fluctuations associated with Parkinson's disease are captured by using an FFT window size of 2048 and a hop length of 512 samples to increase model fidelity. To calculate the magnitude spectrogram [24] $S(f,t)$, use $|X_n(k)|$, where $X_n(k)$ stands for the DFT coefficients. Mathematically, this is represented as follows:

$$S(f,t) = |X_n(k)| \tag{2}$$

  Here, f represents frequency and t represents time.

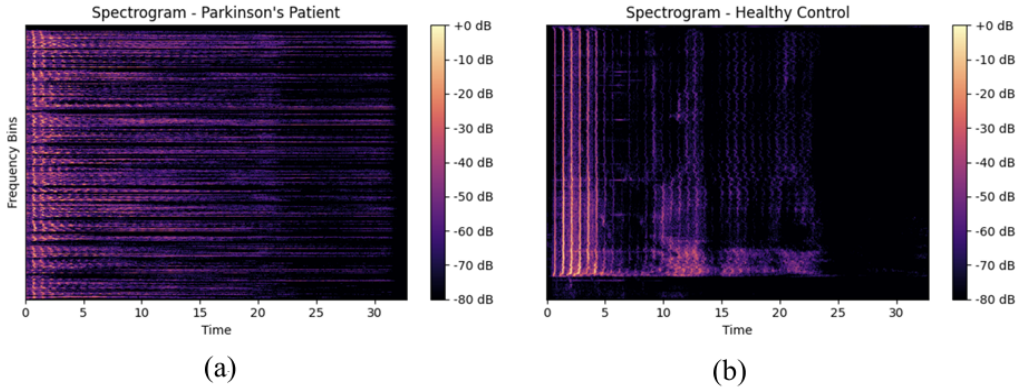


Figure 3. Spectrogram extraction outcomes on the PD dataset using Libros Python 3.13 (**a**) for Parkinson patient and (**b**) for healthy control.

- **Extraction of mel spectrograms**: Parkinson's speech, in which vocal characteristics are altered in ways that are detectable to the ear, is analyzed using mel spectrograms, featured in Figure 4, which remap frequencies to the mel scale in accordance with human hearing. To provide perceptually significant frequency detail, the audio is split into frames, Fourier processed, and converted to 30 mel bands ($n\_melspec = 30$). Equation (3), where S_m is the spectrum magnitude and i is the frequency bin, can be used to get the spectral centroid frequency (FSC) [25]. Here, N represents the total number of bins in the spectrum.

$$FSC = \sum_{N}^{i=1}(iSm(i)) / \sum_{N}^{i=1}(Sm(i)) \tag{3}$$

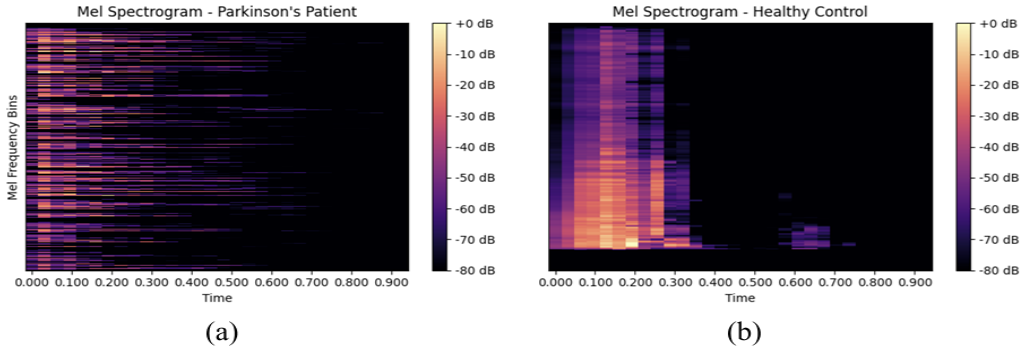


Figure 4. Mel spectrogram extraction outcomes on the PD dataset using Libros Python (**a**) for Parkinson patient and (**b**) for healthy control.

- **Extraction of mel-frequency cepstral coefficients**: The first step in extracting mel-frequency cepstral coefficients (MFCCs) is pre-emphasis, which amplifies higher frequencies and is frequently affected in Parkinson's speech. After being windowed to minimize spectral leakage and framed (20–30 ms), the audio is then FFT-converted to the frequency domain. The Discrete Cosine Transform (DCT) is used to derive 30 MFCCs after applying a mel filterbank on the power spectrum. This procedure [26] can be represented mathematically as follows and represented graphically in Figure 5:

$$MFCCs = DCT(log(Mel\ filterbank(FFT(Windowed\ signal)))) \tag{4}$$

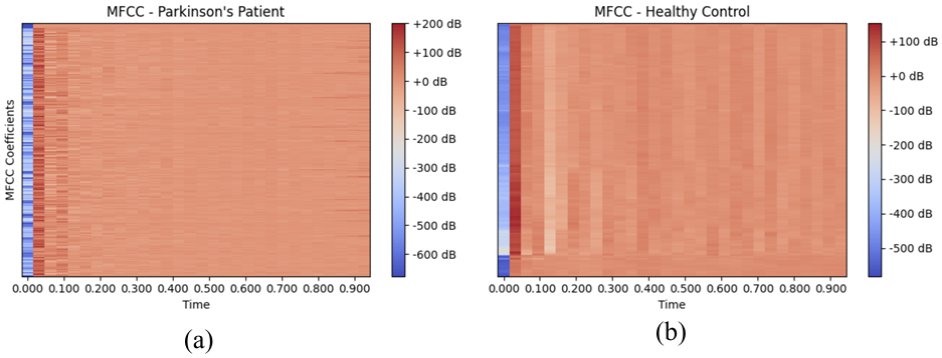


Figure 5. MFCC extraction outcomes on the PD dataset using Libros Python (**a**) for Parkinson's patient and (**b**) for healthy control.

3.3.1. Model Architecture

The proposed approach synthesizes CNNs and a multi-head attention mechanism along with LSTM networks to comprehensively model the spatial and temporal correlations inherent in sequential data, such as audio signals, as depicted in Figure 6 and also as described in Algorithm 1. Initially, the input data undergo processing through CNN layers to extract spatial features and identify local patterns, such as spectrograms or mel-frequency cepstral coefficients (MFCCs). A subsequent pooling layer is employed to reduce dimensionality, preserving essential information while enhancing computational efficiency. The flattened CNN output is then passed through a multi-head attention mechanism, enabling the model to focus on salient segments of the sequence and capture complex relationships through diverse attention heads. Finally, the attention-processed data is fed into LSTM layers with 128 units, which are crucial for modeling long-term dependencies in time series data, particularly in audio-related applications.

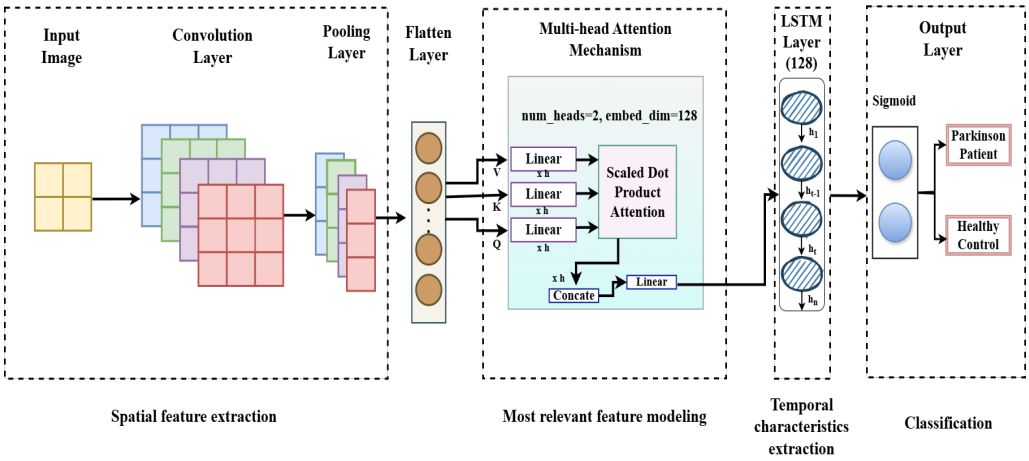


Figure 6. Structure of attention-based hybrid CNN-LSTM model with multi-head attention mechanism for Parkinson identification.

Ultimately, prior to reaching the output layer, the flattened CNN output, attention output, and LSTM output are concatenated into a single feature vector, then processed through dense layers. The classification process predicts whether the input belongs to a healthy control or a Parkinson's patient using a sigmoid activation function. To ensure robustness and mitigate overfitting, regularization techniques such as dropout and L2 regularization are employed, ensuring reliable model performance on unseen data.

The model's detailed structure is shown in Table 2, which also includes specifications for each module. Additionally, the algorithm underlying the proposed approach is provided for comprehensive understanding of the methodology.

Table 2. Network details for the proposed hybrid Parkinson detection model.

Module	Details
Convolutional Layers	Conv1: 64 filters, kernel size = 5, activation = ReLU, kernel_regularizer = L2 (0.01) Conv2: 64 filters, kernel size = 5, activation = ReLU, kernel_regularizer = L2 (0.01)
Max Pooling Layer	MaxPooling1: pool size = 5×5
Dropout Layer	Dropout: 0.5
Convolutional Layers	Conv3: 64 filters, kernel size = 5, activation = ReLU, kernel_regularizer = L2 (0.01) Conv4: 64 filters, kernel size = 5, activation = ReLU, kernel_regularizer = L2 (0.01)
Max Pooling Layer	MaxPooling2: pool size = 5×5
Dropout Layer	Dropout: 0.5
Attention	Multi-head attention, number of heads = 2
Recurrent Layers	LSTM1: 128 units, return sequences = True LSTM2: 128 units, return sequences = False
Dropout Layer	Dropout: 0.5

Algorithm 1 Proposed Model for Parkinson's Disease Detection Using Hybrid CNN–Multi-Head Attention with LSTM

Input: Set of raw input signals $\mathcal{X} = \{x_i(t) \mid i = 1, \ldots, N\}$, each of size $60 \times 94 \times 1$
Output: Predicted labels $\hat{\mathcal{Y}} = \{\hat{y}_i \in \{\text{Parkinson's}, \text{Healthy}\} \mid i = 1, \ldots, N\}$

for each input signal $x_i(t)$ **do**
  Extract mel spectrogram $S_i(t, f)$ and MFCC $C_i(t, m)$:

$$S_i(t, f) = \text{MelSpectrogram}(x_i(t)), \quad C_i(t, m) = \text{MFCC}(x_i(t))$$

  Concatenate $S_i(t, f)$ and $C_i(t, m)$ to form $B_i(t)$:

$$B_i(t) = [S_i(t, f), C_i(t, m)]$$

  Initialize $I_1 = B_i(t)$
  for each convolutional layer $l = 1$ to 4 **do**
    Apply Conv2D with ReLU activation and L_2 regularization
    if l is even **then**
      Apply MaxPooling and Dropout
    end if
    Update I_{l+1} based on layer output
  end for
  Flatten output feature map to $F_{i,\text{flat}}$
  Apply Multi-Head Attention to capture features A_i
  Pass A_i through LSTM layers with dropout (H_i)
  Compute output $\hat{y}_i$ using Sigmoid activation
  Compute binary cross-entropy loss:

$$L(\theta) = -\frac{1}{B} \sum_{i=1}^{B} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

  Update weights using ADAM optimizer
end for

3.3.2. Multi-Head Attention Layer

To enhance the focus on significant spatiotemporal information captured by the CNN, the hybrid model for Parkinson's classification incorporates a multi-head attention mechanism [27] with two attention heads, as illustrated in Figure 7. Following feature extraction by the CNN, the extracted features are utilized as the query (Q), key (K), and value (V) matrices within the attention layer. This configuration enables the model to emphasize relevant patterns across different segments of the input sequence, thereby identifying critical trends within the data.

Subsequently, the attention-enhanced features are processed by LSTM layers, which effectively capture long-term temporal dependencies in the sequence. The final binary classification for Parkinson's disease detection is performed by a dense layer with a sigmoid activation function. The integration of CNNs, multi-head attention, and LSTMs equips the model to effectively capture complex dynamics within sequential audio data, resulting in elevated classification accuracy.

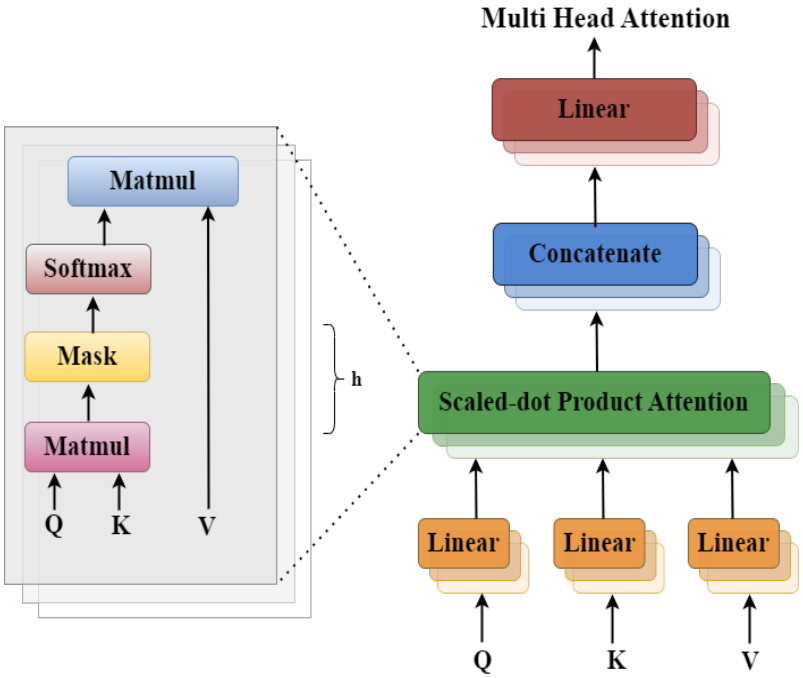


Figure 7. Multi-head attention consists of several attention layers (h) running in parallel.

Let Q, K, and V represent the input queries, keys, and values, respectively. These are linearly projected to the desired embedding space:

$$Q' = Q \cdot (W_Q) \tag{5}$$

$$K' = k \cdot (W_k) \tag{6}$$

$$V' = V \cdot (W_V) \tag{7}$$

$W_Q, W_K,$ and W_V are weight matrices for the projections.

The attention scores are computed as the dot product between the query and key vectors, scaled by the square root of the dimension of the key vectors.

$$\text{Attention Scores} = \frac{Q' \cdot K'^T}{\sqrt{\dim(K')}} \tag{8}$$

Next, softmax is applied to normalize the attention scores across the sequence:

$$\text{Attention Weights} = \text{softmax}\left(\text{Attention Scores}\right) \tag{9}$$

The attended output for each head is obtained by computing the weighted sum of the values using these attention weights:

$$\text{Attended Output} = \text{Attention Weights} \cdot V' \tag{10}$$

3.3.3. Hyperparameter Optimization

The hyperparameters for the model used to classify Parkinson's disease are compiled in Table 3. The Adam optimizer is used to train the model for 30 epochs with a batch size of 32 and a learning rate of 0.001. A dropout rate of 0.5 and L_2 regularization (strength 0.01) are used to avoid overfitting. The output layer employs a sigmoid activation in conjunction with the binary cross-entropy loss function to produce probability scores for the classification of individuals as either healthy controls or Parkinson's patients.

Table 3. Hyperparameter settings for Parkinson disease classification.

Hyperparameters	Value
Epoch	30
Dropout	0.5
Optimizer	Adam
Batch size	32
Learning rate	0.001
Loss function	binary cross-entropy
Regularization kernel	L_2 regularization
Regularizer strength	0.01
Classifier activation function	Sigmoid

4. Experimental Analysis

The trials are conducted on a GPU-supported Google Colaboratory. To handle audio tasks including conversion, modification, and feature extraction, Pydub version 0.25.1 is installed. This package guarantees seamless raw audio pre-processing, facilitating effective data transformation for the classification model of PD. The process of preparing data for model training and evaluation is made more efficient by its interaction with the Python 3.13 environment. Our thorough evaluation of the model's performance with the test set came after 30 epochs of training.

4.1. Results and Discussion

The results section is divided into parts, as initially we will provide ablation study outcomes, followed by the proposed models' performance. Then, a comparison is made between the proposed model and the state-of-the-art model.

4.1.1. Ablative Analysis

Table 4 compares the models that are assessed based on the features that are retrieved and each model's accuracy in classifying Parkinson's disease.

Each model is assessed using a variety of features, such as the spectrogram, mel spectrogram, MFCCs, and a combined Mel_MFCC feature set. Across all architectures, models that use the concatenated Mel_MFCC features consistently attain superior accuracy. Notably, the CNN–multi-head attention with LSTM model reduces the parameter considerably and achieves an outstanding accuracy of 99.00%. Performance is improved by the attention

method, which sharpens the model's focus on important characteristics, particularly when using the Mel_MFCC feature set. Spectrogram (50–52%) and mel spectrogram (79–85%) characteristics shows lower accuracy because they describe the frequency content of the audio signals in a broad way, which may comprise redundant or irrelevant information. These characteristics are less suited to picking up on minute speech abnormalities, such as jitter and pitch changes, which are integral for identifying Parkinson's disease. In contrast, MFCC and Mel_MFCC are drawn to extract speech-specific features, which leads to much higher accuracy. This performance discrepancy emphasizes how crucial feature selection is when creating models for speech-based illness detection.

Table 4. Ablation study of the proposed model according to different extracted features.

Model	Extracted Features	Accuracy (%)
CNN–BiLSTM–multi-head attention	Spectrogram	50.12
	Mel spectrogram	79.90
	MFCC	88.18
	Mel_MFCC	90.22
CNN–BiGRU–multi-head attention	Spectrogram	50.36
	Mel spectrogram	80.94
	MFCC	90.77
	Mel_MFCC	92.73
CNN–GRU–multi-head attention	Spectrogram	50.48
	Mel spectrogram	83.63
	MFCC	92.58
	Mel_MFCC	95.79
CNN–multi-head attention–LSTM (proposed model)	Spectrogram	52.25
	Mel spectrogram	85.58
	MFCC	98.79
	Mel_MFCC	**99.00**

The performance of various models for classifying Parkinson's disease is presented in Table 5 along with their accuracy and parameters (in millions). With 1.05 M parameters, the baseline CNN achieves 97.30% accuracy; however, the CNN-BiLSTM and CNN-GRU models produce inconsistent outcomes. Performance is enhanced by incorporating multi-head attention; with 3.72 M parameters, CNN-LSTM and multi-head attention achieve 99.69% accuracy. The suggested CNN–multi-head attention–LSTM model provides an ideal mix between great performance and reduced complexity, achieving 99.00% accuracy with just 2.70 M parameters. The phrase "parameters" here indicates the total number of each model's trainable and non-trainable parameters. The weights and biases that are refined by back-propagation throughout the training process are known as trainable parameters, and they allow the model to identify patterns in the data. Conversely, non-trainable parameters are set during the training process and can result from batch normalization (e.g., moving mean and variance) or frozen layers in pre-trained models. The computational complexity of each model is specified by the total number of parameters, as shown in Table 5. This function ensures transparency in estimating the complexity and performance of the model by clearly breaking down the parameters that are trainable and those that are not.

Table 5. Evaluated models' performance for Parkinson classification.

Models	Accuracy	Parameters (M)
CNN	97.30	1.05
CNN-BiLSTM	84.27	3.08
CNN-GRU	96.54	3.60
CNN-LSTM	97.59	3.44
CNN–attention–LSTM (soft)	98.39	3.62
CNN–LSTM–attention (soft)	98.20	4.15
Attention before and after LSTM (soft)	98.64	4.15
CNN–attention–LSTM (hard)	90.99	4.35
CNN–attention–GRU (hard)	98.56	4.49
CNN–BiLSTM–multi-head attention	90.22	3.78
CNN–BiGRU–multi-head attention	92.73	4.26
CNN–GRU–multi-head attention	95.79	4.03
CNN-LSTM with multi-head attention	99.69	3.72
CNN–multi-head attention–LSTM (**proposed model**)	99.00	2.70

4.1.2. Proposed Model's Result Analysis

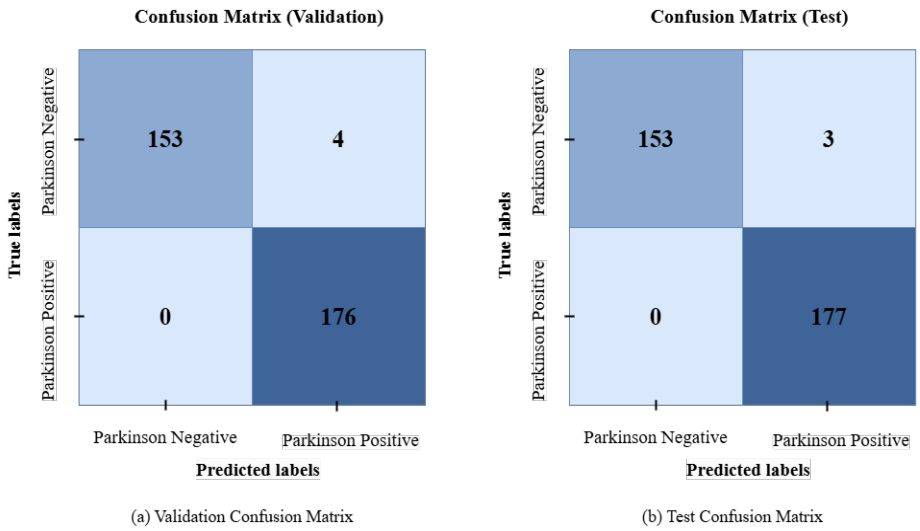
Table 6 summarizes the performance evaluation of the proposed attention-based model for classifying individuals as healthy controls or Parkinson's patients. The model achieves a precision of 1.00 for the Healthy Control class, indicating perfect accuracy in classification, while the recall was 0.98, reflecting that 98% of actual healthy controls were correctly identified. For the Parkinson Patient class, the model demonstrates a precision of 0.98 and a perfect recall of 1.00, successfully detecting all actual patients. Both classes result in an F1-score of 0.99, underscoring the model's robustness and effectiveness in distinguishing between healthy and Parkinson's patients.

As highlighted in the section on related works, the majority of preceding research, such as that conducted by Aditya et al. [15] and Umesh et al. [18], only considers reporting accuracy, which restricts the ability to directly compare precision, recall, and F1-scores. In contrast, our suggested approach provides a thorough assessment, attaining close to flawless precision, recall, and F1-scores for both Parkinson patients and healthy controls.

Table 6. Precision, recall and F1-score of attention-based proposed hybrid model on speech dataset.

Class	Precision	Recall	F1-Score
Healthy Control	1.00	0.98	0.99
Parkinson Patient	0.98	1.00	0.99

The diagnostic efficacy of a binary classification model for Parkinson's disease is demonstrated by the confusion matrices depicted in Figure 8. With only four false positive and no false negatives, the model correctly identified 176 healthy people (true negatives) and 153 Parkinson's patients (true positives) in the validation sample. When the model was used on the test set, it successfully recognized 153 healthy people but was unable to recognize any Parkinson's patients as positive. It produced 3 false negatives and 177 false positives, indicating a sharp decline in performance.



(a) Validation Confusion Matrix (b) Test Confusion Matrix

Figure 8. Confusion matrices for validation and test sets of the proposed hybrid model.

The Receiver Operating Characteristic (ROC) curve shown in Figure 9 illustrates the performance of the evaluated model in classifying individuals as either Parkinson's patients or healthy controls. The orange ROC curve indicates a perfect classification capability, as indicated by an area under the curve (AUC) of 1.00. This result signifies that the model accurately identifies all Parkinson's patients without any false positives or false negatives. The curve approaches the top-left corner of the plot, reflecting a high true positive rate (sensitivity) and a minimal false positive rate across all classification thresholds. Overall, these findings demonstrate that the model exhibits exceptional reliability and accuracy in differentiating between Parkinson's patients and healthy individuals.

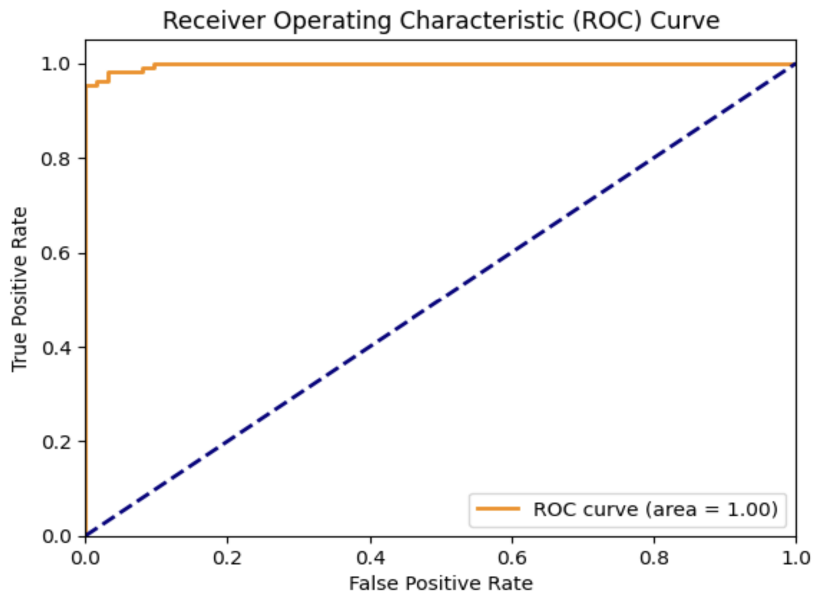


Figure 9. ROC curve demonstrating the model's perfect classification of Parkinson's patients versus healthy controls, with an AUC of 1.00.

4.2. Comparison with Existing Models

A standout performer among the models studied is the proposed hybrid CNN-LSTM model featuring a multi-head attention mechanism; it is noteworthy that there are considerable differences in the experimental setups and datasets utilized in related studies. Our model outperforms methods like the Hybrid CNN-LSTM model by Umesh et al. [18]

(93.49% accuracy) and the NNB method by Aditya et al. [15] (80.88% accuracy), as described in Table 7.

Table 7. Performance comparison of proposed model with existing models and techniques based on accuracy.

Reference	Models/Techniques	Accuracy
Umesh et al. [18]	Hybrid CNN-LSTM	93.49%
Aditya et al. [15]	Nearest Neighbor Boosting (NNB)	80.88%
Gayathri et al. [19]	ADNN ADRNN ADCNN	95.48% 97.29% 98.32%
Raya et al. [6]	KNN SVM MLP RF DT	83.15% 95.26% 98.49% 90.65% 92.77%
Proposed model	CNN–multi-head attention–LSTM	**99.00%**

Some research with lower accuracy can be the result of using fewer feature sets or traditional machine learning algorithms. Our approach uses sophisticated methods like attention mechanism and data augmentation to improve generalization.

Even with its 98.32% accuracy, the ADCNN model by Gayathri et al. [19] falls short of the intended hybrid model. The models of Raya et al. [6] produce a range of results, with the accuracy of their MLP model matching that of the hybrid model. Other strategies, including CNN, SVM, RF, and DT, on the other hand, exhibit a notable performance disparity and significantly lower accuracy.

The proposed hybrid model's outstanding accuracy results from its skillful fusion of CNNs and LSTMs which take advantage of the data's ability to capture both temporal and spatial patterns. Moreover, the addition of a multi-head attention mechanism improves the model's ability to concentrate on important aspects, which adds to its exceptional performance. The incorporation of regularization techniques aids in reducing overfitting, and random oversampling efficiently rectifies class imbalances, all of which contribute to the overall improvement of the model's resilience and predictive power.

The accuracy of the proposed approach is contrasted with that of a number of other models that are utilized in this field which are highlighted in Figure 10. The accuracy of the proposed model is 99.00%, which is higher than the ADCNN model's 98.32% performance. Most models cluster between 90 and 98 percent accuracy, with a range of 80% (NNB) to above 99%. NNM performs poorly at 80%, but ADRNN (97.29%) and CNN-LSTM (93.49%) are high performers. Depending on the application scenario, a notable improvement of percentage points above the best available model is indicated.

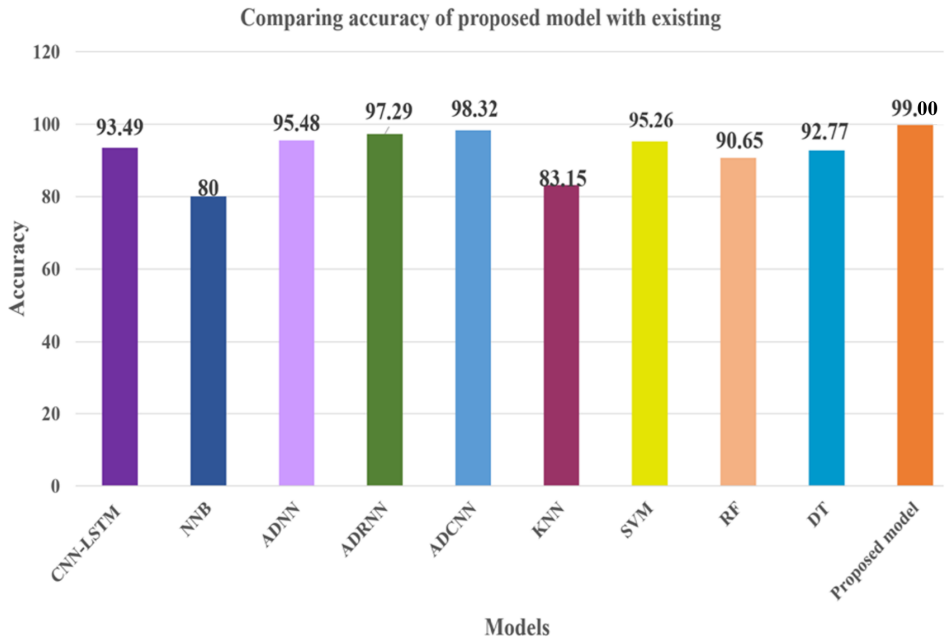


Figure 10. Comparative analysis of the proposed model with existing models.

5. Discussion

This study's remarkable 99.004% classification accuracy demonstrates the promise of speech analysis as a non-intrusive method for early Parkinson's disease identification. Effectively capturing both spatial and temporal patterns in speech data is exhibited by the great performance of the proposed CNN-multi-head attention with LSTM architecture. Nevertheless, despite these positive outcomes, some restrictions and difficulties need to be noted. Even though the dataset consisted of 831 raw audio samples that were supplemented to 3496, a strong performance was guaranteed by the effective application of regularization and augmentation. Furthermore, the study concentrated on speech features; future investigations can be explored with other modalities, such handwriting analysis and gait freezing, to increase the model's resilience. Explainability approaches like Grad-CAM and SHAP will be used in future studies to improve the model transparency and emphasize the most vital characteristics. Although the architecture's complexity guarantees great performance, it also brings attention to the possibility of additional optimization in settings with limited resources. In future, lightweight models and optimization strategies for real-time deployment, like Mobile-Net or quantization, will be investigated. Expanding the dataset to include multi-modal data and enhancing model accessibility will bridge the gap between research and implementation. By tackling these issues and building on the promising results, this study lays the foundation for useful, easily accessible, and expandable Parkinson's disease detection systems.

6. Conclusions

In conclusion, this study highlights the potential of speech analysis as a non-invasive and cost-effective method for the early detection of Parkinson's disease. A hybrid CNN–multi-head attention with LSTM architecture is developed, leveraging the spatial feature extraction capabilities of convolutional networks, the sequential processing strengths of LSTMs, and the attention mechanism to emphasize critical speech features. In view of this, to facilitate the diagnosis of Parkinson's disease through the analysis of vocal characteristics, the study focuses on employing a combination of mel spectrogram and MFCC as spectral features. By utilizing these features, the proposed model achieves a classification accuracy

of 99.00%. This research demonstrates the amazing potential of blending advanced deep learning architectures with conventional feature extraction methods to successfully address real-world medical problems. The work successfully handles the difficulties presented by imbalanced and constrained datasets while simultaneously guaranteeing strong model performance through the use of sophisticated augmentation and pre-processing procedures. These findings underscore the robustness and potential scalability of speech analysis as a reliable and affordable diagnostic tool for remote healthcare applications and early Parkinson's disease screening. Future developments in integrating wearable technologies and real-time deployment could completely transform this field and make Parkinson's disease detection much more convenient and beneficial for international healthcare systems. By connecting technology and clinical practice, this study shows voice analysis as a useful addition to conventional diagnoses, improving access to healthcare worldwide.

Author Contributions: Conceptualization, M.I. and K.A.; methodology, M.I.; formal analysis, M.I.; investigation, M.I. and K.A.; resources, M.I.; writing—original draft preparation, M.I.; writing— review and editing, K.A., M.A.H. and M.A.A.D.; supervision, K.A. and M.A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used for the experiment is publicly available at the following URL: https://ieeexplore.ieee.org/abstract/document/8070308, accessed on 5 April 2024.

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

References

1. Jankovic, J. Parkinson's disease: Clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry* **2008**, *79*, 368–376. [CrossRef]
2. Dorsey, E.R.; Sherer, T.; Okun, M.S.; Bloem, B.R. The Emerging Evidence of the Parkinson Pandemic. *J. Park. Dis.* **2018**, *8*, S3–S8. [CrossRef] [PubMed]
3. Azadi, H.; Akbarzadeh-T, M.R.; Shoeibi, A.; Kobravi, H.R. Evaluating the Effect of Parkinson's Disease on Jitter and Shimmer Speech Features. *Adv. Biomed. Res.* **2021**, *10*, 54._254_21. [CrossRef]
4. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations. *Mov. Disord. Off. J. Mov. Disord. Soc.* **2003**, *18*, 738–750. [CrossRef]
5. Sethi, K. Clinical aspects of parkinson disease. *Curr. Opin. Neurol.* **2022**, *15*, 457–460. [CrossRef]
6. Alshammri, R.; Alharbi, G.; Alharbi, E.; Almubark, I. Machine learning approaches to identify parkinson's disease using voice signal features. *Front. Artif. Intell.* **202**3, *6*, 1084001. [CrossRef] [PubMed]
7. Ziólko, M. Speech Analysis as a Tool for Detection and Monitoring of Medical Conditions: A Review (Preprint). 2020. Available online: https://journals.pan.pl/dlibra/show-content?id=128239 (accessed on 27 January 2025).
8. Jean-Luc, H. Parkinson's disease. *Rev. Prat.* **2022**, *5510*, 1129–1134.
9. Armstrong, M.J.; Okun, M.S. Diagnosis and treatment of parkinson disease: A review. *JAMA* **2020**, *323*, 548–560. [CrossRef] [PubMed]
10. Wu, H.; Soraghan, J.J.; Lowit, A.; Caterina, G.D. A deep learning method for pathological voice detection using convolutional deep belief networks. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; International Speech Communication Association (ISCA): Hyderabad, India, 2018.
11. Sakar, B.E.; Serbes, G.; Sakar, C.O. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PLoS ONE* **2017**, *12*, e0182428.
12. Rahman, W.; Lee, S.; Islam, M.S.; Antony, V.; Ratnu, H.; Ali, M.R.; Mamun, A.A.; Wagner, E.; Jensen-Roberts, S.; Little, M.A.; et al. Detecting Parkinson's disease using a web-based speech task: Observational study. *J. Med. Internet Res.* **2021**, *23*, e26305. [CrossRef]
13. Marino, S.; Ciurleo, R.; Lorenzo, G.D.; Barresi, M.; Salvo, S.D.; Giacoppo, S.; Bramanti, A.; Lanzafame, P.; Bramanti, P. Magnetic resonance imaging markers for early diagnosis of parkinson's diseaseff. *Neural Regen. Res.* **2012**, *7*, 611–619. [PubMed]

14. Maffia, M.; Micco, R.D.; Pettorino, M.; Siciliano, M.; Tessitore, A.; Meo, A.D. Speech rhythm variation in early-stage Parkinson's disease: A study on different speaking tasks. *Front. Psychol.* **2021**, *12*, 668291. [CrossRef] [PubMed]
15. Shastry, K.A. An ensemble nearest neighbor boosting technique for prediction of Parkinson's disease. *Healthc. Anal.* **2023**, *3*, 100181. [CrossRef]
16. Hossain, M.A.; Amenta, F. Machine learning-based classification of parkinson's disease patients using speech biomarkers. *J. Parkinson's Dis.* **2023**, *14*, 95–109. [CrossRef] [PubMed]
17. Govindu, A.; Palwe, S. Early detection of parkinson's disease using machine learning. *Procedia Comput. Sci.* **2023**, *218*, 249–261. [CrossRef]
18. Lilhore, U.K.; Dalal, S.; Faujdar, N.; Margala, M.; Chakrabarti, P.; Chakrabarti, T.; Simaiya, S.; Kumar, P.; Thangaraju, P.; Velmurugan, H. Hybrid cnn-lstm model with efficient hyperparameter tuning for prediction of parkinson's disease. *Sci. Rep.* **2023**, *13*, 14605. [CrossRef]
19. Nagasubramanian, G.; Sankayya, M. Multi-variate vocal data analysis for detection of parkinson disease using deep learning. *Neural Comput. Appl.* **2020**, *33*, 4849–4864. [CrossRef]
20. Dimauro, G.; Nicola, V.D.; Bevilacqua, V.; Caivano, D.; Girardi, F. Assessment of speech intelligibility in parkinson's disease using a speech-to-text system. *IEEE Access* **2017**, *5*, 22199–22208. [CrossRef]
21. Abayomi-Alli, O.O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; Misra, S. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics* **2022**, *11*, 3795. [CrossRef]
22. Surina, S. Time and Pitch Scaling in Audio Processing. Available online: https://www.surina.net/article/time-and-pitch-scaling.html (accessed on 27 January 2025).
23. Wongvorachan, T.; He, S.; Bulut, O. A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information* **2023**, *14*, 54. [CrossRef]
24. Oppenheim, A.V.; Schafer, R.W. *Discrete-Time Signal Processing*, 3rd ed.; Pearson: Upper Saddle River, NJ, USA, 2009.
25. Müller, M. *Fundamentals of Audio and Music Processing: With Applications to Signal Processing and Music Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2015.
26. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed.; Pearson: London, UK, 2020.
27. Deora, P.; Ghaderi, R.; Taheri, H.; Thrampoulidis, C. On the optimization and generalization of multi-head attention. *arXiv* **2023**, arXiv:2310.12680.