ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2023.2.202

A Robust Voice Pathology Detection System Based on the Combined BiLSTM–CNN Architecture

Rimah Amami[1,✉], Rim Amami[2], Chiraz Trabelsi[3], Sherin Hassan Mabrouk[4], Hassan A. Khalil[5]

[1]Computer Department, Deanship of Preparatory Year and Supporting Studies, Imam AbdulRahman bin Faisal University, Dammam, KSA
[2]Basic Science Department, Deanship of Preparatory Year and Supporting Studies, Imam AbdulRahman bin Faisal University, Dammam, KSA
[3]Institut Montpelliérain Alexander Grothendieck, UMR CNRS 5149, Place Eugéne Bataillon, 34090, Montpellier, France
Département Sciences et Technologies, Centre Universitaire de Mayotte, 3 Route Nationale, 97660, Dembéni, France
[4]Self-Development Department, Deanship of Preparatory Year and Supporting Studies, Imam AbdulRahman bin Faisal University, Dammam, KSA
[5]Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt
raamami@iau.edu.sa[✉],rabamami@iau.edu.sa,chiraz.trabelsi@univ-mayotte.fr,shmabrouk@iau.edu.sa,H.a.khalil@zu.edu.eg

Abstract

Voice recognition systems have become increasingly important in recent years due to the growing need for more efficient and intuitive human-machine interfaces. The use of Hybrid LSTM networks and deep learning has been very successful in improving speech detection systems. The aim of this paper is to develop a novel approach for the detection of voice pathologies using a hybrid deep learning model that combines the Bidirectional Long Short-Term Memory (BiLSTM) and the Convolutional Neural Network (CNN) architectures. The proposed model uses a combination of temporal and spectral features extracted from speech signals to detect the different types of voice pathologies. The performance of the proposed detection model is evaluated on a publicly available dataset of speech signals from individuals with various voice pathologies(MEEI database). The experimental results showed that the hybrid BiLSTM-CNN model outperforms several classifiers by achieving an accuracy of 98.86%. The proposed model has the potential to assist health care professionals in the accurate diagnosis and treatment of voice pathologies, and improving the quality of life for affected individuals.

Keywords: *Voice Pathology Detection, Convolutional Neural Network, BiLSTM, Hybrid Systems, MEEI Voice Disorders Database.*

Received: 10 September 2023
Accepted: 30 October 2023
Online: 05 November 2023
Published: 20 December 2023

1 Introduction

Speech is an essential means of communication that allows humans to convey their thoughts and emotions. However, for individuals with voice disorders, this fundamental human capability is compromised. Early detection and diagnosis of voice disorders can lead to effective treatment and management, thereby improving the quality of life of affected individuals. Furthermore, voice pathology is a common disorder that affects a large number of individuals worldwide, and it can be caused by various factors such as vocal cord paralysis, laryngitis, and tumors. Indeed, voice pathologies can severely impact the quality of life of affected individuals and timely diagnosis is essential for effective treatment.

The accurate and timely detection of voice pathology is crucial for effective treatment and management of the condition. In recent years, deep learning techniques have shown great promise in improving the accuracy and efficiency of voice pathology detection systems.

In recent years, the use of machine learning (ML) and deep learning (DL) techniques has gained popularity in the field of voice pathology detection. Among these techniques, convolutional neural networks (CNNs) [19, 13, 23] and long short-term memory LSTM [20, 7, 11, 14] networks have demonstrated remarkable performance in speech processing tasks.

In this research paper, we propose a novel hybrid architecture that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for voice pathology detection. Our proposed architecture leverages the strengths of both CNNs and LSTMs to achieve improved accuracy and robustness in detecting voice pathology.

The CNNs are used for feature extraction, which helps to identify the discriminative features in the input speech signals. The LSTM networks are then used to capture the temporal dynamics of the speech signals and to model the long-term dependencies between the input features.

It is noted that CNNs are well-known for their abil-

ity to extract spatial features from images, and they have also been applied successfully in speech recognition tasks. In contrast, LSTMs are powerful tools for modeling sequential data, and they have shown great promise in speech processing applications. By combining these two architectures, we can exploit the spatial and temporal information present in voice signals to improve the accuracy of voice pathology detection.

The hybrid bidirectional LSTM (BiLSTM) networks and deep learning can be used for voice pathology detection, which involves identifying and diagnosing voice disorders such as vocal nodules, polyps, and laryngitis. An hybrid BiLSTM network is a type of neural network that combines the Bidirectional Long Short-Term Memory (BiLSTM) architecture with other deep learning techniques, such as Convolutional Neural Networks (CNNs) [16, 18] and/or Fully Connected Layers (FCLs). This type of network can be particularly useful for processing sequential data, such as speech signals, as it can capture long-term dependencies in the data. To use hybrid LSTM networks for voice pathology detection, one approach is to preprocess speech signals using Mel-frequency cepstral coefficients (MFCCs) or other feature extraction techniques. These features can then be fed into the hybrid LSTM network, which can learn to classify them into different voice pathology categories. Training an hybrid BiLSTM network for voice pathology detection requires a large dataset of labeled speech signals, which can be obtained from patients with known voice pathologies. The network can then be trained using a supervised learning approach, where the network learns to classify the input features based on the corresponding pathology labels. Once trained, the hybrid LSTM network can be used to predict the presence of voice pathologies in new speech signals, allowing for early diagnosis and intervention. This approach has the potential to improve the accuracy and efficiency of voice pathology diagnosis, leading to better treatment outcomes for patients.

Our proposed hybrid architecture is trained and evaluated using a publicly available dataset of voice recordings from patients with different types of voice disorders. We compare the performance of our hybrid architecture with that of traditional CNN and LSTM models, as well as with other state-of-the-art approaches for voice pathology detection.

The results of our experiments demonstrate that our proposed hybrid architecture achieves superior performance compared to traditional CNN and LSTM models and outperforms other state-of-the-art approaches in terms of accuracy, robustness, and generalization ability. Our findings suggest that the proposed hybrid architecture holds great potential for improving the accuracy and efficiency of voice pathology detection systems, and can ultimately aid in the timely and accurate diagnosis and treatment of this common disorder.

The remainder of this paper is organized as follows. Section 2 describe the voice pathologies detection systems and the main challenge to be faced. Section 3 provides an overview of the related work in the field of voice pathology detection. Section 4 illustartes the proposed hybrid architecture in detail, including the CNN and LSTM components. Section 5 presents the experimental setup and results. Finally, section 5 concludes the paper and discusses the potential applications and future directions of the proposed approach

2 Voice Pathologies Detection Systems

Voice pathology detection systems are computer-based technologies that are designed to assist in the diagnosis and treatment of speech disorders. These systems use advanced algorithms to analyze vocal characteristics such as pitch, tone, and articulation in order to detect and diagnose voice disorders.

Voice pathology detection systems can be useful in a variety of settings, including speech therapy, medical diagnosis, and research. They can help clinicians and researchers to identify and track changes in vocal characteristics over time, which can be valuable in developing effective treatment plans.

There are several different types of voice pathology detection systems, ranging from simple software applications to more advanced machine learning models. Some of the most common types of systems include:

Acoustic analysis systems: These systems analyze voice recordings to detect abnormalities in pitch, tone, and other vocal characteristics. They can be used to diagnose a wide range of voice disorders, including dysphonia, vocal cord paralysis, and laryngeal cancer.

Machine learning models: These systems use advanced algorithms to analyze large datasets of voice recordings in order to detect patterns and identify potential voice disorders. They are particularly useful in research settings where large amounts of data can be analyzed quickly and efficiently.

Speech recognition systems: These systems use natural language processing and machine learning algorithms to analyze speech patterns and detect abnormalities in speech production. They can be used to diagnose speech disorders such as stuttering, apraxia, and dysarthria.

Overall, voice pathology detection systems have the potential to significantly improve the diagnosis and treatment of speech disorders. They offer a non-invasive and cost-effective way to analyze vocal characteristics, and can be used in a variety of clinical and research settings.

2.1 Importance of Voice Pathologies Detection Systems

Voice pathologies detection systems are important because they can help identify and diagnose various voice disorders, such as dysphonia, vocal fold nodules, laryngitis, and other conditions that affect the voice. These systems use advanced technology to analyze and quantify various aspects of a person's voice, including pitch,

loudness, voice quality, and other parameters, to detect signs of pathology or abnormalities.

Early detection and diagnosis of voice disorders are essential for effective treatment and management. Left untreated, these conditions can lead to more severe health problems, including voice fatigue, chronic hoarseness, and even vocal fold damage. In some cases, voice disorders can also impact a person's quality of life, leading to social anxiety, communication difficulties, and reduced self-confidence.

Voice pathology detection systems can provide accurate and objective assessments of a person's voice, helping healthcare professionals diagnose and treat voice disorders more effectively. By using these systems, doctors and speech-language pathologists can monitor changes in a person's voice over time, track treatment progress, and adjust treatment plans accordingly.

Overall, voice pathology detection systems are crucial for improving the accuracy and efficiency of voice disorder diagnosis and treatment, ultimately leading to better outcomes for patients

2.2 Challenges and Limitations of Voice Pathologies Detection Systems

voice pathology detection systems are designed to identify and diagnose voice disorders based on speech samples recorded by the user. While these systems have the potential to improve the accuracy and efficiency of voice disorder diagnosis, there are several challenges and limitations associated with their use. Some of these include:

Limited availability of high-quality speech data: Voice pathology detection systems rely on large datasets of high-quality speech samples to train their algorithms. However, obtaining such data can be challenging, particularly for rare disorders or for populations that are not well-represented in existing datasets.

Variability in speech patterns: Speech patterns can vary widely between individuals, even those with the same disorder. This can make it difficult for voice pathology detection systems to accurately diagnose disorders based on speech samples alone.

Limited diagnostic capabilities: While voice pathology detection systems can be helpful in identifying certain voice disorders, they may not be able to diagnose all types of disorders or identify the underlying causes of a disorder.

Need for specialized equipment: Many voice pathology detection systems require specialized equipment, such as a microphone or speech analysis software, which can be expensive and may not be readily available in all settings.

Lack of user acceptance: Some individuals may be hesitant to use voice pathology detection systems, either due to concerns about privacy or because they prefer to receive a diagnosis from a human healthcare provider.

Overall, while voice pathology detection systems have the potential to improve the accuracy and efficiency of voice disorder diagnosis, they are not without limitations. Researchers and developers must continue to work to address these challenges in order to make these systems more effective and accessible to individuals who could benefit from them.

3 Literature Review

The state of the art in hybrid LSTM networks and deep learning for voice recognition has advanced significantly in recent years, thanks to advances in hardware, software, and machine learning algorithms.

Hybrid LSTM networks are a type of neural network that combines the long short-term memory (LSTM) architecture with other neural network architectures, such as convolutional neural networks (CNNs), to improve the accuracy and robustness of voice recognition systems. LSTM networks are particularly effective for modeling sequences of data, such as speech signals, because they can capture long-term dependencies and context information.

One of the most widely used deep learning techniques for voice recognition is the deep neural network (DNN). DNNs are trained on large datasets of speech signals and can learn to extract useful features from the input data. These features are then used to classify the input signal into one of several predefined categories, such as speech or non-speech.

Another important deep learning technique for voice recognition is the convolutional neural network (CNN). CNNs are commonly used for image classification tasks, but they can also be used for speech recognition. By applying convolutional filters to the input speech signal, a CNN can learn to identify important features, such as phonemes and prosody, that are useful for speech recognition.

Alex Graves and Navdeep Jaitly in cite [8] proposes an hybrid LSTM network that combines deep bidirectional LSTM layers with a hidden Markov model to improve speech recognition accuracy.

In [9], the authors presents an end-to-end system for speech recognition that uses a deep LSTM network with a CTC loss function to directly map input speech to output text.

On the other hand, the study in [2] extends the work of the previous paper to include an hybrid LSTM network with convolutional layers and residual connections, achieving state-of-the-art results on several speech recognition benchmarks.

The proposed study in [21] introduce an hybrid LSTM network that combines a deep LSTM acoustic model with a separate LSTM language model to improve speech recognition accuracy.

G.Saon and al. in [22] explores the use of deep LSTM networks for speaker-independent speech recognition, achieving state-of-the-art results on several benchmarks. Overall, these research papers demonstrate the effectiveness of hybrid LSTM networks and deep

learning for voice recognition tasks, and highlight the importance of integrating language models and other techniques to improve accuracy.

Recent advances in deep learning for voice recognition also include the use of attention mechanisms, which allow the network to focus on specific parts of the input signal that are most relevant for the task at hand. This has led to significant improvements in accuracy, especially for speech recognition in noisy environments.

Overall, the state of the art in hybrid LSTM networks and deep learning for voice recognition is rapidly advancing, and these techniques are expected to play an increasingly important role in a wide range of applications, including virtual assistants, speech-to-text systems, and voice-controlled devices.

Furthermore, A. Graves and al. proposes in [10] a deep recurrent neural network (RNN) model, including long short-term memory (LSTM) units, for speech recognition. The model achieves state-of-the-art performance on the TIMIT dataset.

The study of Hannun and al. [12] introduces a deep neural network model, including convolutional and LSTM layers, for end-to-end speech recognition. The model achieves state-of-the-art performance on the Switchboard and CallHome datasets.

Chorowski and al. [4] introduces an attention-based model, including an LSTM encoder and decoder, for speech recognition. The model dynamically focuses on different parts of the input sequence and achieves state-of-the-art performance on the LibriSpeech dataset.

A generative model is presented in [17], based on deep convolutional and dilated convolutional layers, for speech synthesis. The model achieves state-of-the-art performance on the Blizzard Challenge dataset.

All these papers showed the effectiveness of hybrid LSTM networks and deep learning for voice recognition tasks, and they provided valuable insights for designing and improving speech recognition models.

There have been several recent research papers that have explored the use of hybrid LSTM networks and deep learning for voice pathology detection. Here are some related works: In [5], a deep learning approach is introduced to detect pathological voice disorders and a LSTM autoencoder hybrid with multi-task learning solution is proposed with spectrogram as input feature.

A bidirectional long short-term memory (BI-LSTM) is used to classify different types of pathological voices in [3]. The proposed system achieved an accuracy of 92.7% in classifying pathological voice.

Ksibi and al. in [15], propose a speech pathology identification system based onthe convolutional network integrated with the recurrent neural network (RNN) with LSTM layers.

Overall, these studies demonstrate the potential of hybrid LSTM networks and deep learning for voice pathology detection, and suggest that these techniques could be useful in developing accurate and efficient diagnostic tools for voice disorders.

4 Proposed Architecture

4.1 Description of the Proposed Hybrid Approach

The Hybrid Bidirectional LSTM networks and deep learning have been very successful in improving speech recognition systems.

BiLSTM networks are a type of recurrent neural network (RNN) that can learn long-term dependencies in sequential data, making them well-suited for speech recognition tasks where the context of a spoken word can greatly impact its interpretation. However, BiLSTMs alone may not be able to capture all the complexities of speech, which is why hybrid architectures have been developed.

In an hybrid BiLSTM network, multiple layers of BiLSTMs are stacked on top of each other, with each layer responsible for learning different features of the speech signal. The output of the BiLSTM layers is then passed through one or more fully connected layers to make the final classification decision (see Figure 1).

Deep learning techniques, such as convolutional neural networks (CNNs), can also be used in conjunction with BiLSTM networks to further improve the accuracy of speech recognition. CNNs are particularly useful for extracting local features from the audio signal, such as the frequency and amplitude of individual phonemes. These features can then be fed into the BiLSTM network to capture long-term dependencies.

One popular architecture for speech recognition using hybrid LSTM and deep learning is tfhe Listen, Attend and Spell (LAS) model. This model uses a CNN to extract local features from the audio signal, followed by an attention mechanism to focus on relevant parts of the signal during the decoding process. The output of the attention mechanism is then fed into a stacked BiLSTM network to make the final prediction.

BiLSTM is commonly used in deep learning applications for voice recognition, including the identification of voice pathologies. Voice pathologies can refer to any number of conditions that affect a person's ability to speak or communicate effectively, such as vocal cord nodules, polyps, or paralysis.

One way that BiLSTM networks can be used for voice pathology detection is by training the model on large datasets of audio recordings of healthy and pathological voices. The model would learn to recognize patterns in the audio data that are indicative of different voice pathologies. For example, it might learn to distinguish between healthy and pathological vocal cords based on variations in pitch, tone, or other acoustic features.

In order to train a BiLSTM model for voice pathology detection, a large dataset of labeled audio recordings is necessary. This dataset should include both healthy and pathological voices, with a sufficient number of samples for each pathology of interest. The model can then be trained using supervised learning techniques, where it learns to classify the recordings based on the presence or absence of different voice

pathologies.

4.2 Architecture of Hybrid BiLSTM-CNNs for Voice Pathology Detection

The voice pathologies, such as hoarseness, breathiness, and strain, can have a significant impact on a person's quality of life. Accurately identifying and diagnosing these pathologies can be challenging, as they can be subtle and difficult to detect.

The proposed type of network can process sequential data, such as audio waveforms, and extract meaningful features from them. In the proposed model based on an hybrid BiLSTM network and CNNs for voice pathologies detection, the different layers in the network have specific roles in processing the input audio data and extracting relevant features for classification.

Here is a brief overview of the role of each layer:

- **Input Layer:** The input layer receives the audio waveform data and converts it into a format that can be processed by the network.
- **Convolutional Layer:** The convolutional layer applies a set of filters to the input audio signal to extract features that are relevant for classification. The filters are learned through training the network.
- **BiLSTM Layer:** The Bidirectional Long Short-Term Memory (BiLSTM) layer processes the output of the convolutional layer and extracts temporal information that is important for voice pathology detection. LSTMs are a type of recurrent neural network (RNN) that can effectively capture long-term dependencies in time-series data. The main difference between the traditional LSTM and BiLSTM is that BiLSTM adds one more LSTM layer that reverses the direction of information flow.
- **Fully Connected Layer:** The fully connected layer receives the output of the BiLSTM layer and performs classification based on the learned features. This layer is typically followed by a softmax layer that produces class probabilities.
- **Output layer:** This layer would provide the final prediction and could use a softmax activation function to produce a probability distribution over the possible classes.

The combination of these layers in an hybrid BiLSTM network enables the network to effectively capture both spatial and temporal features of the input audio data for accurate identification of voice pathologies. The network is trained using a large dataset of labeled audio recordings of healthy and pathological voices from the MEEI database, allowing it to learn to recognize patterns in the data that are indicative of different types of voice disorders.

Figure 1 illustrates the architecture of the proposed model based on BiLSTM-CNN.

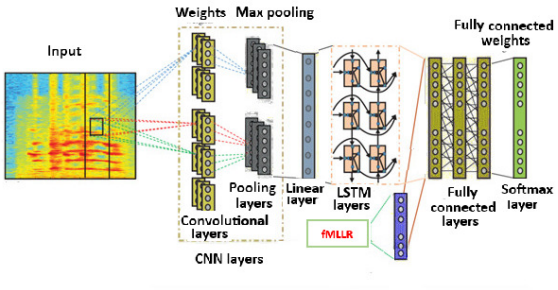

Figure 1: The proposed system architecture.

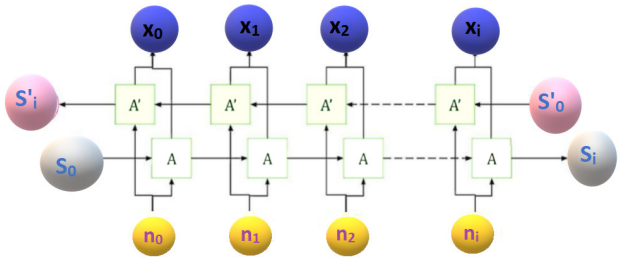

Figure 2: Bidirectional LSTM layer Architecture.

To start with, a few convolutional layers are used to decrease the frequency variance present in the input signal. Specifically, two convolutional layers, each with 256 feature maps, are initially utilized due to the small feature dimension for speech. After passing through these two convolutional layers, the feature map is reduced to a smaller size of around 16, eliminating the need for further locality modeling and invariance removal. The authors in [1] have suggested that a 99 frequency-time filter for the first convolutional layer and a 43 frequency-time filter for the second convolutional layer are enough to cover the entire frequency-time space, hence these filter sizes are employed for the first and second convolutional layers, respectively.

In our model, we begin with max pooling using a pooling size of 2 a for both layers. To reduce the dimensionality of the output without compromising accuracy, a linear layer is applied following the CNN layers, resulting in 256 outputs.

Next, the output of the frequency modeling is fed into a BLSTM layer which models the signal in time. Two BLSTM layers and three FC layers are used, with each BLSTM layer having 820 cells and a 512 unit projection layer for dimensionality reduction.

The Bidirectional LSTM(BiLSTM) is based on two LSTM layers; the first layer is to process the input in the forward direction while the second layer is to process in the backward direction. Hence, the BiLSTM model, consider the two directions forward and backward to process the data in order to better mapping the data. Figure 2 illustrates the architecture of a Bidirectional LSTM.

It must be pointed out that n_i denotes the input samples, the token output is denoted by x_i. Further-

more, the LSTM nodes are described by A and A'. Indeed, the output of x_i is the fusion of the LSTM nodes.

Then, the output is passed through fully connected (FC) layers to generate higher-order feature representations that are easily distinguishable between different classes. Each FC layer contains 1024 hidden units. To account for variability in speech due to differences in speakers' accent, loudness, etc., we use the fMLLR technique, which cannot be directly modeled by CNNs. Previously, fMLLR transformation was applied to log-mel features.

The suitable number of layers in an hybrid BiLSTM network and CNNs [1] for voice pathology detection depends on several factors, including the complexity of the data, the size of the dataset, and the computational resources available. In general, a deeper network can capture more complex patterns and features in the data, which can lead to better performance in terms of accuracy. However, deeper networks also require more computational resources and can be more prone to overfitting, especially when the dataset is small. For the voice pathology detection, typical architecture for an hybrid BiLSTM network might include several BiLSTM layers followed by several fully connected layers. As a starting point, we could consider a network with 2-3 BiLSTM layers and 2-3 fully connected layers. It's important to note that the optimal number of layers and architecture can vary depending on the specific task and dataset, so it is important to perform experimentation and tuning to find the best architecture for your particular use case.

In order to investigate the best performance of the proposed architecture, several structures are tested by varying the number of the BiLSTM layers and FC layers. The accuracy of voice pathology detection is based on the number of BiLSTM layers and FC layers, as shown in Table 1. The results demonstrates that when the number of BiLSTM layers is up to 3, it enhance the overall performance.

Table 1: Investigation of the suitable Number of BiLSTM and Fully connected layers for the proposed system.

Number of BiLSTM	fully connected layers	EER
1 BiLSTM	4 fully connected layers	07.7
2 BiLSTM	3 fully connected layers	07.2
3 BiLSTM	3 fully connected layers	05.3
3 BiLSTM	2 fully connected layers	04.1
4 BiLSTM	2 fully connected layers	03.3
5 BiLSTM	2 fully connected layers	04.6

4.3 Datasets

Healthy voices are characterized by clear and consistent speech patterns. They typically exhibit good control of pitch, volume, and tone. Healthy voices are also free from vocal abnormalities, such as hoarseness, breathiness, or strain [6].

Unhealthy voices, on the other hand, may exhibit a range of speech disorders and voice abnormalities. These can include hoarseness, which is characterized by a rough or scratchy voice, breathiness, which is characterized by a weak or airy voice, and strain, which is characterized by a strained or forced voice. Unhealthy voices may also exhibit pitch breaks, tremors, or other fluctuations in pitch or volume.

It is important to note that a person's voice can be affected by a variety of factors, including illness, injury, stress, and environmental factors such as air pollution or excessive vocal use. A trained professional, such as a speech-language pathologist or otolaryngologist, can provide a more detailed evaluation of a person's voice and make recommendations for treatment or therapy. There are several datasets used for voice pathology detection, including:

1. *MEEI:* The Massachusetts Eye and Ear Infirmary (MEEI) dataset contains recordings from 80 patients, including individuals with normal voice and those with various voice disorders, such as nodules, polyps, and paralysis.
2. *Saarbrücken Voice Database:* The Saarbrücken Voice Database (SVD) includes recordings from 200 healthy individuals and 198 patients with various voice disorders, including hoarseness, breathiness, and strain.
3. *GRBAS:* The Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) scale is a perceptual evaluation tool used to assess voice quality. The GRBAS dataset includes audio recordings from 120 individuals, including those with normal voice and those with various voice disorders.
4. *KayPENTAX Database:* The KayPENTAX Database includes audio recordings from 120 individuals with various voice disorders, including nodules, polyps, and tumors.

In this paper, we used MEEI database to test the robustness of the proposed hybrid BiLMST-CNN for voice patologies detection [24, 25].

In order to investigate the performance of the proposed BiLSTM-CNNs architecture for the voice pathology detection, the MEEI database was tested. It contains 53 healthy samples and 724 samples with voice disorders. The speech datasets were capture with a rate of 25kHz or 50kHz and 16 bits of resolution. The healthy voices were recorded for 3 seconds and the unhealthy voices were recorded for 1 second. For the experimental sets, the duration of each frame is fixed to 30ms and an Hamming window was used to extract the speech frames. In this study, we used 53 healthy voices and 200 pathological voices (Keratosis/Vocal Poly/Adductor/Paralysis). Table 2 illustrates the MEEI datasets used in this study and the different disorders:

Table 2: Healthy and unhealthy voices samples from the MEEI database.

Disorder	Male	Female
Healthy voices		
	21	32
Pathological voices		
Paralysis	38	42
Keratosis	21	19
Vocal Polyp	21	18
Adductor	15	26

5 Results and Discussion

The detection of normal and the different pathological voices rates are shown in table 3 using the hybird BiLSTM-CNN architecture. The detection rates are compared using the most popular acoustic features: Mel-frequency cepstral coefficients (MFCC) and Perceptual Linear Prediction coefficients (PLP). In this study, the detection rates revealed that MFCC outperformed PLP and reaches higher precision because MFCC is able to extract useful cepstral features from the voice signal.MFCC features are observed to be better in keep the voice specific features and characteristic.. The performance of our system is proved using both BiLSTM-CNN and MFCC features. Furthermore, the experimental results generated by the proposed method to detect voice pathology detection, are in overall very promising.
In order to conduct the experimental test and investigate the effectiveness of the proposed method, several evaluation metrics were applied such as: Equal Error Rate(ERR), Detection Cost Function (DCF), Sensitivity and Specificity. The formulae for those evaluation metrics are as follows:

$$\mathbf{EER} = \frac{(FAR + FRR)}{2},$$

where

$$\mathbf{FAR} = \frac{FP}{(FP + TN)}$$

and

$$\mathbf{FRR} = \frac{FN}{(TP + FN)}$$

$$\mathbf{Sensitivity} = \frac{TP}{TP + FN}$$

$$\mathbf{Specificity} = \frac{TN}{TN + FP}$$

$$DCF = \sum_{t=1}^{L} DCF(\alpha_t) = \sum_{t=1}^{L}\sum_{y=1}^{N} \pi_y C(\alpha_t|\theta_y) P_e(\alpha_t|\theta_y)$$

where $\mathbf{DCF}(\alpha_t) = \sum_{y=1}^{N} \pi_y C(\alpha_t|\theta_y) P_e rr(\alpha_t|\theta_y)$

It must be pointed out that α denotes a taking action and $P_e(\theta)$ is the error probability of the detection system of the class θ. Indeed, the error $P_e(\alpha_t|\theta_y)$ depends on the action and the correct class of samples.

Table 4 shows the performance of the proposed hybrid BiLMST-CNN for voice patologies detection compared to different systems using classifiers such as DNN, DeepSVM and SVM. Those experimental results demonstrated the effectiveness of the proposed system that achieved an accuracy of 98.86% compared to 91.33% with DNN and 93.89% with DeepSVM.

The voice pathologies detection rate achieved based on the hybrid BiLSTM-CNN is of the order of 98.86%. This result outperformed the rates generated by using MFCCs features. Then, we investigated the robustness of the proposed method by using different classifier such as Deep Neural Network (DNN), DeepSVM and One-Vs-One Support Vector Machines (SVM). We reach an improvement of 4.84% with the DeepSVM classifier compared to the SVM classifier. Meanwhile, the detection with the DeepSVM classifier outperforms the detection using DNN by 2%.

The experimental results in 4 stresses the efficiency of the proposed hybrid BiLSTM-CNN in order to recognized for normal and pathological voices. Overall, the hybrid BiLSTM network combined with deep learning provides accurate and efficient voice pathology detection, which can be useful in diagnosing and treating voice disorders.

6 Conclusion

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) that is commonly used in deep learning applications for voice recognition, including the identification of voice pathologies.
In this study, a pre-trained CNN is used to extract features from the normal and pathological voices and then input these features into their hybrid BiLSTM network.
The BiLSTM networks are employed to capture the temporal dynamics of the speech signals.

The obtained experimental results demonstrated the effectiveness of an hybrid BiLSTM-CNN architecture for the detection of voice pathologies. The high detection rates achieved by the proposed method suggests that it could be a valuable tool for diagnosing voice pathologies in clinical settings.

As a future work, we suggest use different multi-pathologies detectors to investigate the robustness of the proposed system and to consider the possibility to find out the level of voice pathology.

7 Data Availability

The datasets generated during and/or analysed during the current study are available online here: https://colab.research.google.com/drive/1Yb9EKj0uTtNiky-nz3ANRtiaVKBUkQdu

Table 3: Comparison of EER(%), Efficiency (DCF(%)), Sensitivity(%) and Specificity(%) for the different voice pathology detection systems using different Features and the proposed hybrid BiLSTM-CNN Architecture.

System	Disorder	EER	DCF	Sensitivity	Specificity
12 PLP	**Normal**	11.22±04.65	87.13±03.21	86.12	87.90
	Edema	09.07±03.33	89.04±2.07	85.30	89.29
	Paralysis	10.20±03.35	89.06±01.96	88.37	89.00
	Keratosis	11.05±02.04	88.19±02.66	85.91	87.33
	Vocal Poly	10.13±03.08	89.22±02.51	86.20	88.14
	Adductor	09.22±02.37	91.97±01.18	89.11	89.41
12 MFCC	**Normal**	01.02±00.53	98.73±00.81	99.02	99.27
	Edema	01.74±00.63	99.44±00.81	98.79	99.01
	Paralysis	01.66±00.59	99.34±00.83	98.89	98.94
	Keratosis	01.03±00.60	98.68±00.89	98.97	98.90
	Vocal Poly	01.15±00.23	99.30±00.72	98.66	99.00
	Adductor	00.87±00.17	99.02±00.51	99.13	98.98

Table 4: Detection Rates of the proposed system BiLSTM-CNN and different systems, such as DNN, DeepSVM and SVM Based Classifier.

Overall detection %			
BiLSTM-CNN	**DNN**	**DeepSVM**	**SVM**
98.86	91.33	93.89	89.06

References

[1] Amami, R., Al Saif, S. A., Amami, R., Eleraky, H. A., Melouli, F., and Baazaoui, M. The use of an incremental learning algorithm for diagnosing covid-19 from chest x-ray images. *MENDEL 28*, 1 (2022), 1–7.

[2] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (2016), PMLR, pp. 173–182.

[3] AnilKumar, V., and Reddy, R. V. S. Classification of voice pathology using different features and bi-lstm. In *2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)* (2023), IEEE, pp. 1–4.

[4] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. Attention-based models for speech recognition. *Advances in neural information processing systems 28* (2015).

[5] Dávid Sztahó, K. G., and Gábriel, T. M. Deep learning solution for pathological voice detection using lstm-based autoencoder hybrid with multi-task learning. In *I14th International Joint Conference on Biomedical Engineering Systems and Technologies* (2021), pp. 135–141.

[6] Fu, D., Zhang, X., Chen, D., and Hu, W. Pathological voice detection based on phase reconstitution and convolutional neural network. *Journal of Voice* (2022).

[7] Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. Learning precise timing with lstm recurrent networks. *Journal of machine learning research 3*, Aug (2002), 115–143.

[8] Graves, A., and Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning* (2014), PMLR, pp. 1764–1772.

[9] Graves, A., Jaitly, N., and Mohamed, A.-r. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding* (2013), IEEE, pp. 273–278.

[10] Graves, A., Mohamed, A.-r., and Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (2013), Ieee, pp. 6645–6649.

[11] Graves, A., and Schmidhuber, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks 18*, 5-6 (2005), 602–610.

[12] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

[13] Hema, C., and Marquez, F. P. G. Emotional speech recognition using cnn and deep learning techniques. *Applied Acoustics 211* (2023), 109492.

[14] Kim, M. H., Kim, J. H., Lee, K., and Gim, G.-Y. The prediction of covid-19 using lstm algorithms. *International Journal of Networked and Distributed Computing 9*, 1 (2021), 19–24.

[15] Ksibi, A., Hakami, N. A., Alturki, N., Asiri, M. M., Zakariah, M., and Ayadi, M. Voice pathology detection using a two-level classifier based on combined cnn–rnn architecture. *Sustainability 15*, 4 (2023), 3204.

[16] Minh, H. T., Anh, T. P., et al. A novel lightweight dcnn model for classifying plant diseases on internet of things edge devices. *MENDEL 28*, 2 (2022), 41–48.

[17] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[18] Parak, R., and Juricek, M. Intelligent sampling of anterior human nasal swabs using a collaborative robotic arm. *MENDEL 28*, 1 (2022), 32–40.

[19] Pittala, R. B., Tejopriya, B., and Pala, E. Study of speech recognition using cnn. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (2022), IEEE, pp. 150–155.

[20] Rather, A. M. Lstm-based deep learning model for stock prediction and predictive optimization model. *EURO Journal on Decision Processes 9* (2021), 100001.

[21] Sak, H., Senior, A., and Beaufays, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Interspeech 2014* (2014).

[22] Saon, G., Soltau, H., Emami, A., and Picheny, M. Unfolded recurrent neural networks for speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association* (2014).

[23] Schuler, J. P. S., Romani, S., Abdel-Nasser, M., Rashwan, H., and Puig, D. Color-aware two-branch dcnn for efficient plant disease classification. *MENDEL 28*, 1 (2022), 55–62.

[24] Souli, S., Amami, R., Soltani, A., and Yahia, S. B. On the use of deep learning and scattering transform for pathological voices recognition. In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)* (2022), vol. 1, IEEE, pp. 1055–1058.

[25] Souli, S., Amami, R., and Yahia, S. B. A robust pathological voices recognition system based on dcnn and scattering transform. *Applied Acoustics 177* (2021), 107854.