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.

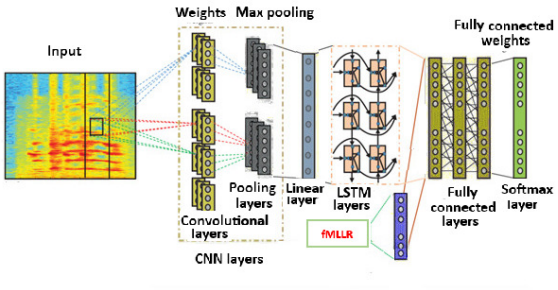

Figure 1: The proposed system architecture.

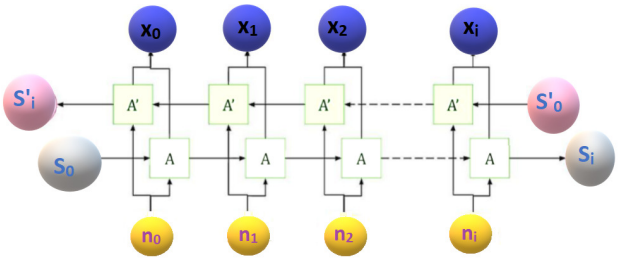

Figure 2: Bidirectional LSTM layer Architecture.

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.