اولین کنفرانس بین المللی هوش مصنوعی

1st INTERNATIONAL CONFERENCE ON

Artificial Intelligence

۷ تا ۹ اسفند ماه ۱۴۰۳

ICAI

First International Conference on Artificial Intelligence

Efficient DL Models for Voice Pathology Detection in Healthcare Applications using Sustained Vowels

مدل های یادگیری عمیق کارآمد برای تشخیص آسیب شناسی گفتار در پزشکی با استفاده از آوای واکه های پایدار

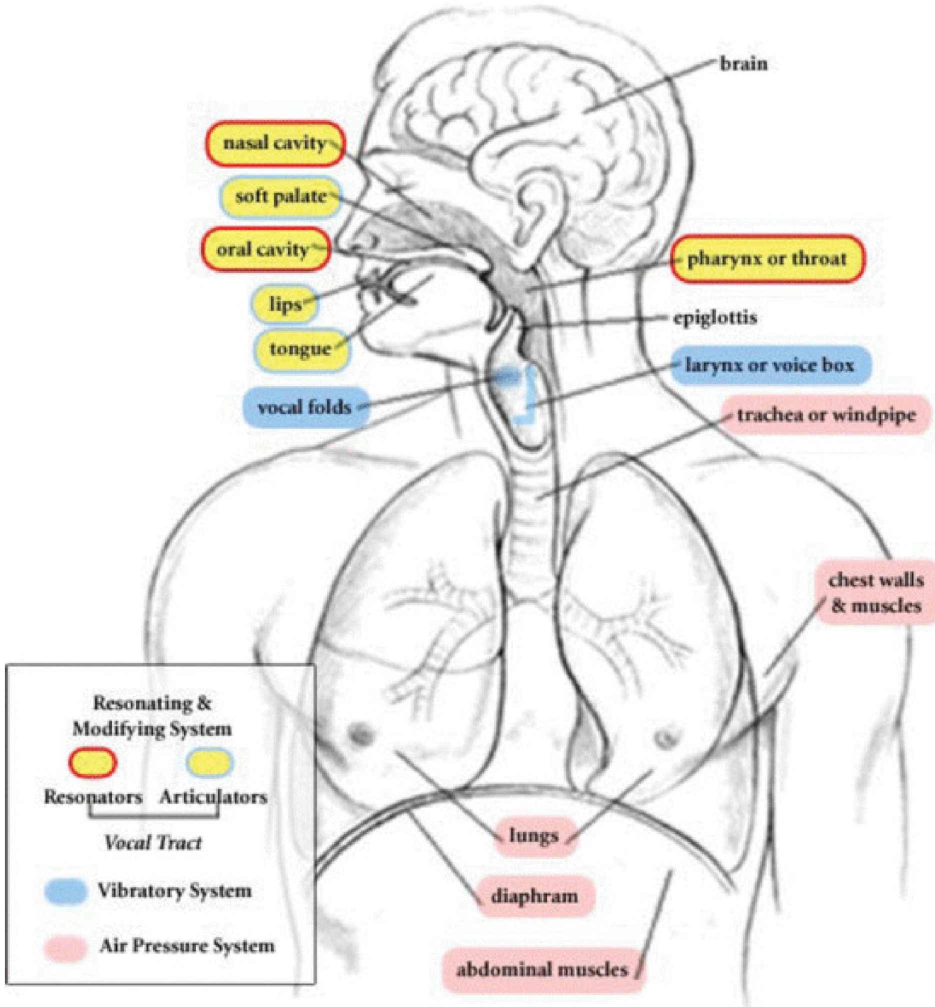
Sahar Farazi, & Yasser Shekofteh

Faculty of Computer Science and Engineering (CSE),

Shahid Beheshti University (SBU), Iran

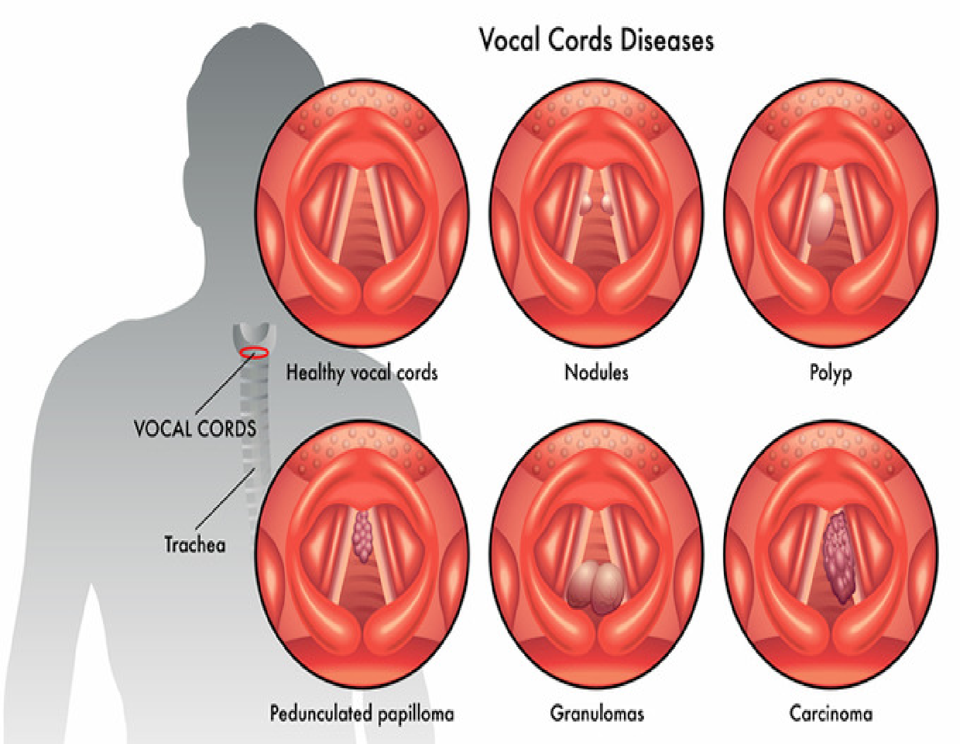
دانشگاه شهید بهشتی

Voice Production System:



- The power source for voice production is the airstream which originates in the lungs and is supported by a diaphragm. The voice production system shown in Figure begins at the vocal cords and terminates at the mouth. The vocal tract includes the larynx, pharynx above it, mouth, and the nasal cavity.

Voice Disorders:



- According to the American Speech-Language-Hearing Association, “Voice disorders occur when voice quality, pitch, and loudness differ or are inappropriate for an individual’s age, gender, cultural background, or geographic location.

- 110% rise in speech disorders among children (ages 0-12) post-pandemic.
- 1 in 5 Americans report voice disorders due to voice tech and occupational use.
- 18% of elderly (60+) suffer from voice-related disorders.

Why Automatic voice Pathology Detection?

AI-Powered Voice Pathology Detection: Key Benefits

□ Early & Accurate Detection

AI analyzes subtle acoustic features, improving diagnosis sensitivity & specificity.

💰 Non-Invasive & Cost-Effective

Eliminates the need for invasive tests; requires only a microphone & software.

⚡ Automation & Efficiency

Processes large voice data quickly, reducing workload for healthcare professionals.

Remote & Telemedicine Applications

Enables screening via smartphones, benefiting underserved regions.

📊 Continuous Monitoring & Personalized Treatment

Tracks voice changes over time for better therapy adjustments.

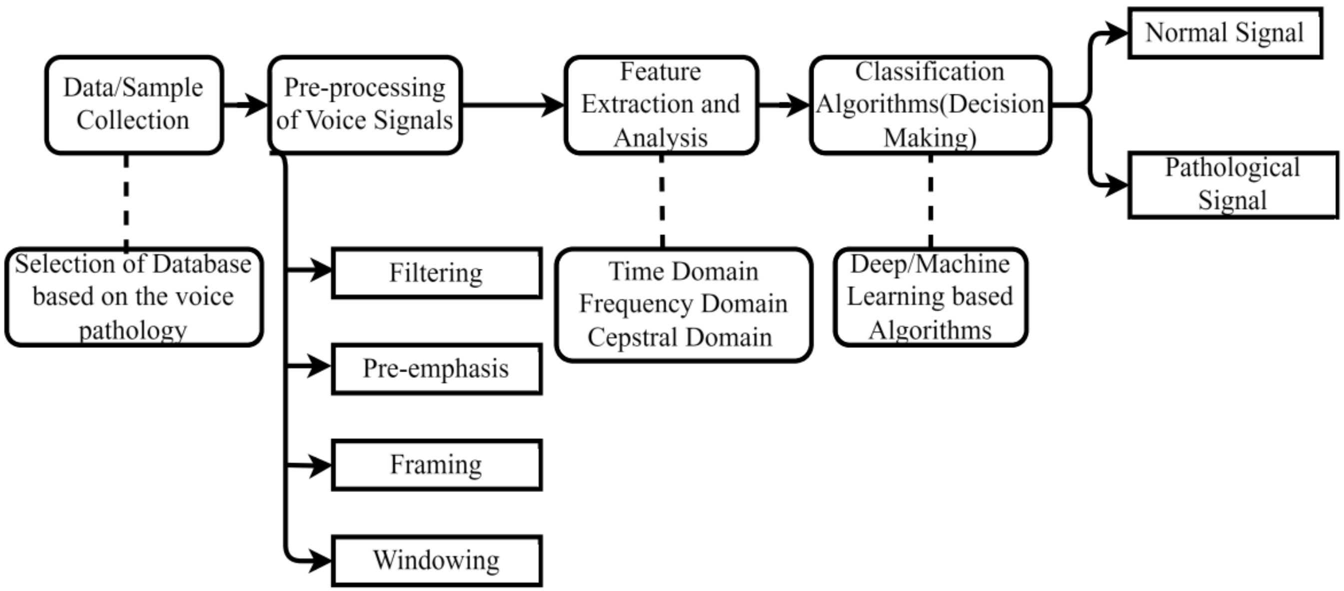
🔍 Detecting Subtle Patterns in Disorders

Identifies complex voice issues like vocal cord paralysis, Parkinson's, and ALS.

📡 Integration with Emerging Technologies

Works with IoT & wearables for real-time monitoring & speech therapy.

Automatic Voice Pathology Detection System:



Related Works:

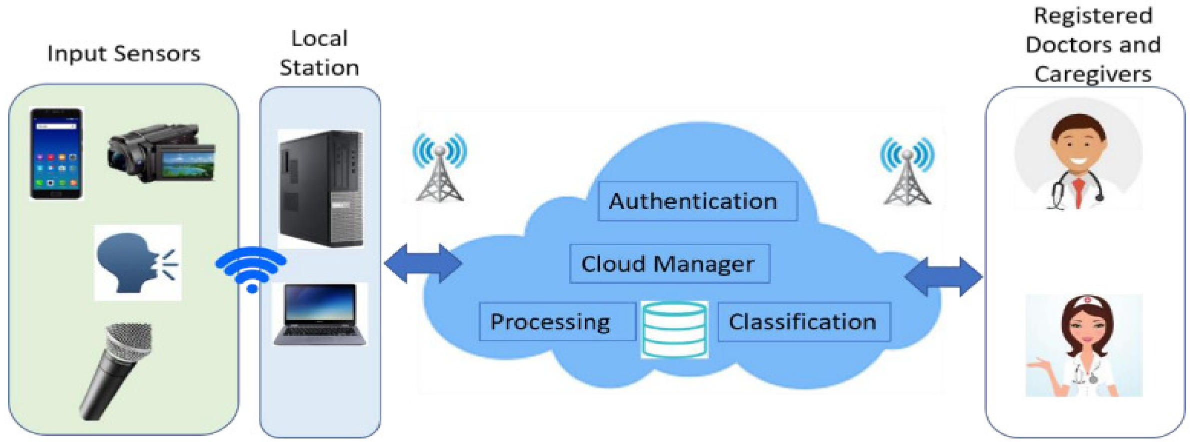
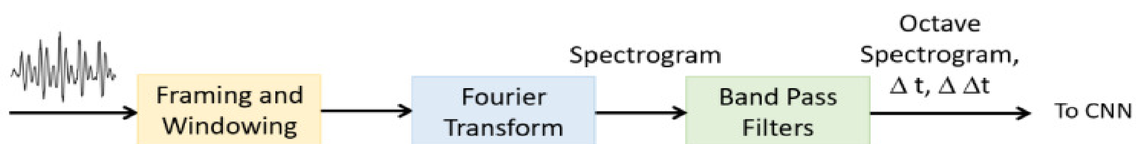


Figure 2. Mobile healthcare framework.



- ✓ A VPD system was developed within a mobile healthcare framework.
- ✓ Smart devices were utilized to capture and process voice signals, leveraging transfer learning with CNN models such as *VGG-16* and *CaffeNet*.
- ✓ Using the *Saarbrücken Voice Disorder* (SVD) database.
- ✓ The system achieved an accuracy of 97.5%, emphasizing the potential of mobile platforms in improving voice pathology diagnostics

- **Dataset:** AVFAD (University of Aveiro, Portugal).
- **Participants:** 709 total (346 with vocal pathologies, 363 healthy).
- **Recording Types:** Sustained vowels (/a/, /u/, /i/) – 3 repetitions each. Reading predefined text & six sentences. Spontaneous speech samples.
- **Sampling Rate:** 48 kHz for all recordings.

Waveform of a healthy and an unhealthy sample for vowel /a/:

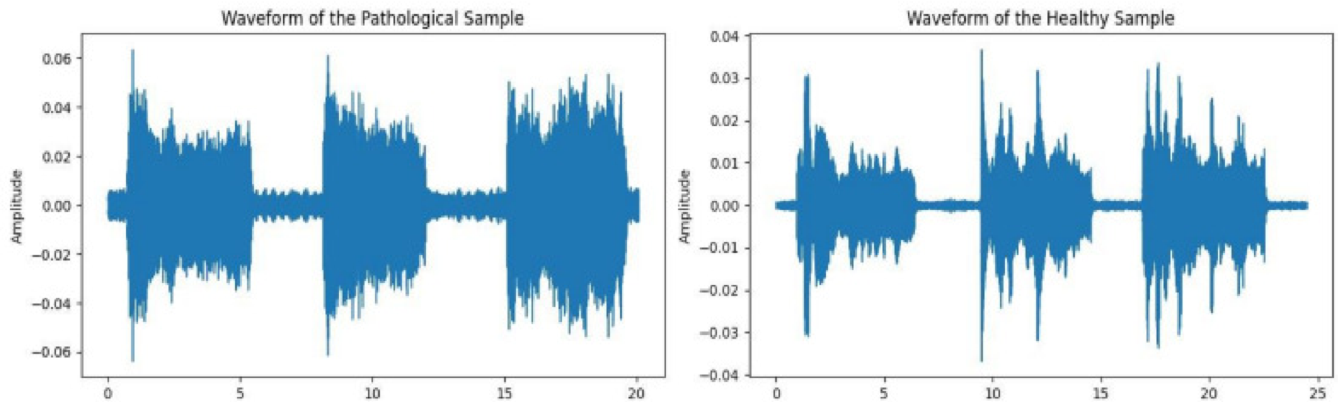


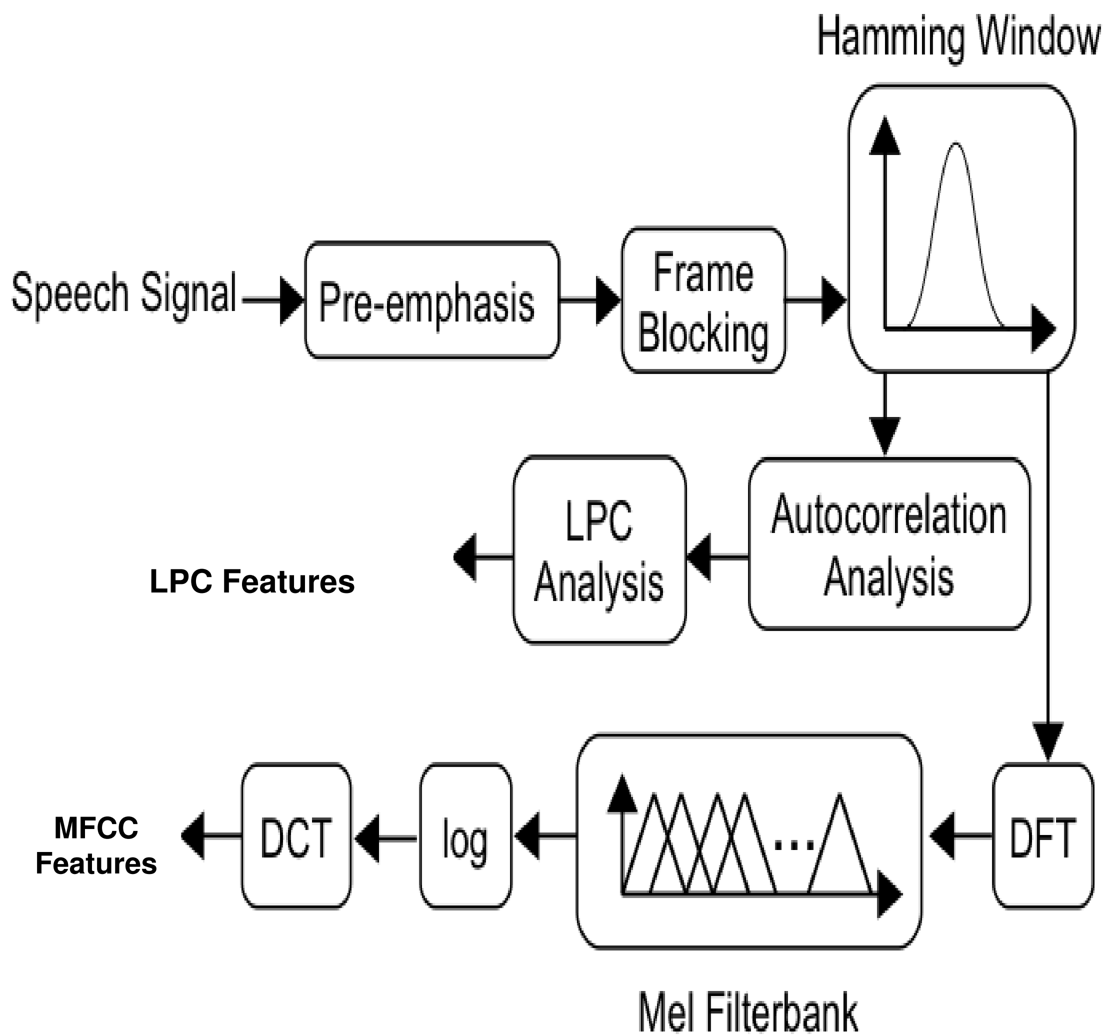
Table 1. Data Splitting Methodology in This Study for the AVFAD Dataset

Data	Train		Test		Validation	
Gender	***Male***	***Female***	***Male***	***Female***	***Male***	***Female***
Normal	73	162	22	50	18	37
Pathologic	64	161	20	49	13	37
Total (Gender)	137	323	42	99	31	74
Total (All)	460		141		105	

Table 2. Duration (sec) statistics for the vowels /a/, /i/, and /u/.

Vowels	Min	Max	Mean	Mean + STD
/a/	3.81	110.92	14.61	21.81
/i/	3.81	121.32	14.81	22.34
/u/	3.63	344.51	14.55	29.09

Feature extraction:



- **MFCC** mimics human auditory perception, effectively representing **timbre and spectral shape**. It is particularly useful in identifying **subtle frequency changes** caused by voice disorders.
- **LPC** models the vocal tract's resonant properties, making it excellent for capturing **speech production characteristics** and detecting **abnormal vocal cord vibrations**.

The Based CNN Model:

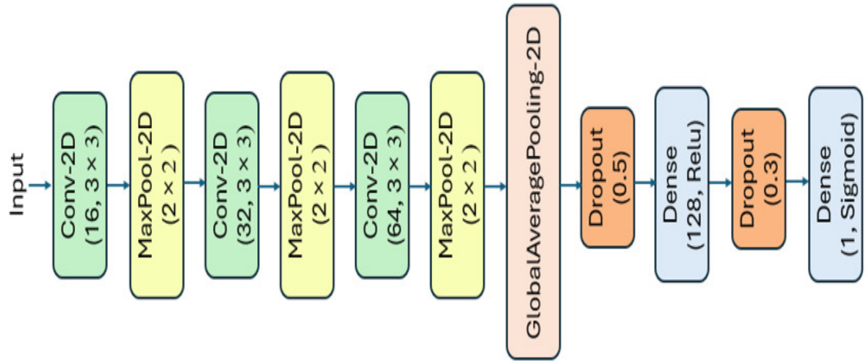


Table 3: The CNN model parameters and details.

Input Layer	Input Shape (Number of Frames, Feature Size, 1)
Convolutional Layer 1	Kernel Size: 3 x 3, Filters: 16, Activation: RELU, Padding: Same
Max pooling Layer 1	Pool Size: 2 x 2
Convolutional Layer 2	Kernel Size: 3 x 3, Filters: 32, Activation: RELU, Padding: Same
Max pooling Layer 2	Pool Size: 2 x 2
Convolutional Layer 3	Kernel Size: 3 x 3, Filters: 64, Activation: RELU, Padding: Same
Max pooling Layer 3	Pool Size: 2 x 2
Global Average Pooling	2 Dimensional
Dropout Layer 1	Rate: 0.5
Dense Layer 1	Units: 128, Activation: RELU
Dropout Layer 2	Rate: 0.3
Dense Layer 2	Units: 1, Activation: Sigmoid

Results:

Table 4: LPC-Based CNN Model - Validation & Test Accuracies for Different Vowels

Vowel type	Frames (Mean+STD) with silence	Frames (Mean+STD) without Silence	First 15 Seconds (with silence)
Vowel /i/	Valid:0.8952 **Test: 0.8591**	Valid:0.8571 **Test: 0.8098**	Valid:0.8857 Test: 0.8380
Vowel /a/	Valid:0.8666 Test: 0.8239	Valid:0.8476 Test: 0.7676	Valid:0.9142 **Test: 0.8450**
Vowel /u/	Valid:0.7619 Test: 0.8028	Valid:0.7333 Test: 0.7183	Valid:0.7809 Test: 0.7816

Table 5: MFCC-Based CNN Model - Validation & Test Accuracies for Different Vowels

Vowel type	Frames (Mean+STD) with silence	Frames (Mean+STD) Without Silence	First 15 Seconds (with silence)
Vowel /i/	Valid:0.8761 **Test:0.8661**	Valid:0.8476 **Test:0.8239**	Valid:0.8666 **Test:0.8521**
Vowel /a/	Valid:0.8857 Test:0.8309	Valid:0.8761 Test:0.7535	Valid:0.8761 Test:0.8380
Vowel /u/	Valid:0.9238 Test:0.8591	Valid:0.9047 Test:0.8098	Valid:0.9047 Test:0.8309

- **LPC-Based CNN Model**

•**Best Test Accuracy: 0.8591** (Vowel **/i/**, Frames with silence).

•**Best Test Accuracy without Silence: 0.8098** (Vowel **/i/**).

•**Best Test Accuracy for First 15 Seconds: 0.8450** (Vowel **/i/**).

- **MFCC-Based CNN Model**

•**Best Test Accuracy: 0.8661** (Vowel **/i/**, Frames with silence).

•**Best Test Accuracy without Silence: 0.8239** (Vowel **/i/**).

•**Best Test Accuracy for First 15 Seconds: 0.8521** (Vowel **/i/**).

- **Overall Best Accuracy**

•**MFCC-Based CNN Model performed best with vowel /i/** (Test Accuracy **0.8661**).

- **Base CNN Model 1**
  - **Architecture:** Three convolutional layers with 16, 32, and 64 filters, respectively, followed by max-pooling max-pooling layers.
  - **First Dense layer neurons:** 128.
  - **Parameters:** 31,745 (~124 KB)
  - **Validation Accuracy:** 0.8761
  - **Test Accuracy:** 0.8661

- **Small CNN Model 2**
  - **Architecture:** Two convolutional layers with 8 and 16 filters, followed by max-pooling layers. The third convolutional layer and its max-pooling layer were discarded.
  - **First Dense layer neurons:** 64.
  - **Parameters:** 1,721 (~6.72 KB)
  - **Validation Accuracy:** 0.8380
  - **Test Accuracy:** 0.7464

- **Small CNN Model 3**
  - **Architecture:** Two convolutional layers with 16 and 32 filters, followed by max-pooling layers. The third layers. The third convolutional layer and its max- and its max-pooling layer were discarded. discarded.
  - **First Dense layer neurons:** 64.
  - **Parameters:** 6,977 (~27.25 KB)
  - **Validation Accuracy:** 0.8571
  - **Test Accuracy:** 0.8309
- **Small CNN Model 4**
  - **Architecture:** Three convolutional layers with 8, 16, and 32 filters, each followed by max-pooling layers.
  - **First Dense layer neurons:** 64.
  - **Parameters:** 8,065 (~31.5 KB)
  - **Validation Accuracy:** 0.8666
  - **Test Accuracy:** 0.8521

Best Result:

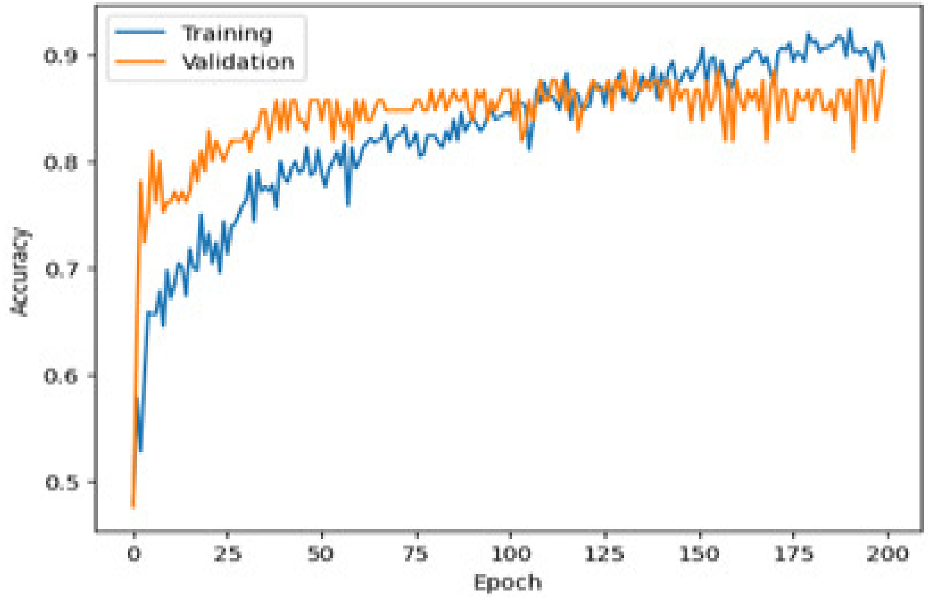


Figure 1: Validation Accuracy Curve for Optimal CNN Model on 20 MFCCs (Vowel /i/)

Table 6: Precision, Recall, and F1-Score for Healthy (0) and Unhealthy (1) Samples of 20 MFCCs for Vowel /i/ Using the Optimal CNN Model

Label	Precision	Recall	F1-Score
0	0.95	0.81	0.87
1	0.83	0.96	0.89

ICAI
First International Conference on Artificial Intelligence

Future Works:

- **Enhancing LPC Features:** Further exploration for deeper insights.
- **Optimizing Feature Extraction:** Varying coefficient numbers to improve classification.
- **Combining Vowel Sounds:** Using multiple vowels for better accuracy.
- **Leveraging Pre-Trained Models:** Enhancing efficiency in pathology detection.
- **CNN Model Improvements:** Exploring different **kernel sizes** for speech signals and testing **GELU activation** instead of ReLU.

References:

1. Abdulmajeed, N.Q., et. al., *A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions*, Journal of Intelligent Systems, 2022. **31**(1): p. 855-875.

2. Latif, S., et al., *Speech technology for healthcare: Opportunities, challenges, and state of the art*, IEEE Reviews in Biomedical Engineering, 2020. **14**: p. 342-356.

3. Alhussein, M. and G. Muhammad, *Voice pathology detection using deep learning on mobile healthcare framework.* IEEE Access, 2018. **6**: p. 41034-41041.

4. Muhammad, G. and M. Alhussein, *Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection.* IEEE Access, 2021. **9**: p. 89198-89209.

5. Hossain, M.S., G. Muhammad, and A. Alamri, *Smart healthcare monitoring: a voice pathology detection paradigm for smart cities.* Multimedia Systems, 2019. **25**(5): p. 565-575.

6. Farazi, S., and Shekofteh, Y., *Voice pathology detection on spontaneous speech data using deep learning models*, International Journal of Speech Technology, 2024. **27**(3), p. 739–751.

7. Syed, S.A., et. al., *Comparative Analysis of CNN and RNN for Voice Pathology Detection,* BioMed Research International, 2021. **2021**(1), p. 1–8.

8. Jesus, L.M., et al., *The advanced voice function assessment databases (AVFAD): Tools for voice clinicians and speech research*, in *Advances in Speech-language Pathology*. 2017, IntechOpen.

9. Xie, X., et. al., *A Voice Disease Detection Method Based on MFCCs and Shallow CNN,* Journal of Voice, 2023. In press.

10. Harar, P., et al. Voice pathology detection using deep learning: a preliminary study. in 2017 international conference and workshop on bioinspired intelligence (IWOBI). 2017. IEEE.

11. Pützer, M. and W. Barry, Saarbrücken Voice Database, Institute of Phonetics, Saarland University. 2009.

12. Mesallam, T.A., et al., *Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms,* Journal of healthcare engineering, 2017. **2017**(1): p. 8783751.

13. Mohammed M.A., *et al.*, "Voice pathology detection and classification using convolutional neural network model," *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10113723.

14. Muhammad, G. *Voice pathology detection using vocal tract area*. in *2013 European Modelling Symposium*. 2013. IEEE.

15. Oliveira, B. F., et. al., Combined sustained vowels improve the performance of the Haar wavelet for pathological voice characterization. In 2020 IWSSIP, pp: 381-386, IEEE.

16. Ribas, D., et. al., On the Problem of Data Availability in Automatic Voice Disorder Detection, In HEALTHINF, 2023. pp. 330–337.

17. Sidhu, M.S., et. al., *MFCC in audio signal processing for voice disorder: a review*. Multimedia Tools and Applications, 2024, In press, p. 1-21.

18. Li, Z., et al., *A survey of convolutional neural networks: analysis, applications, and prospects.* IEEE transactions on neural networks and learning systems, 2021. **33**(12): p. 6999-7019.

19. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting*. The journal of machine learning research, 2014. **15**(1): p. 1929-1958.

20. Firooz, S., et al., *Improvement of automatic speech recognition systems utilizing 2D adaptive wavelet transformation applied to recurrence plot of speech trajectories*. Signal, Image and Video Processing, 2024. **18**(2): p. 1959-1967.

21. Shekofteh, Y. *What can phone attractors in RPS tell us? A study of dynamic information in speech signals for phone classification purposes*, Applied Acoustics, 2023. **211**: p. 109534.

22. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv.org*, Apr. 10, 2015. https://arxiv.org/abs/1409.1556.

23. S. Farazi and Y. Shekofteh, “Evaluation of phone posterior probabilities for pathology detection in speech data using deep learning models,” *International Journal of Speech Technology*, Jan. 2025, doi: 10.1007/s10772-024-10166-w.

Thank you

Sahar Farazi
Farazisahar75@gmail.com