Synthetic Data for AI-Driven Detection of Laryngeal Cancer

Beenish Zia[1]*, Farshid Taghizadeh[2]

[1]Intel Corporation, Hillsboro, Oregon, USA

[2]Oregon Health Science University (OHSU), Portland, Oregon, USA

Research Article

Received: 09-Apr-2025, Manuscript No. GRCS-25-164206; **Editor assigned**: 11-Apr-2025, Pre QC No. GRCS-25-164206 (PQ); **Reviewed:** 25-Apr-2025, QC No. GRCS-25-164206; **Revised:** 02-May-2025, Manuscript No. GRCS-25-164206 (R); **Published:** 30-June-2025, DOI: 10.4172/2229-371X.16.2.003

***For Correspondence:**
Beenish Zia, Intel Corporation, Hillsboro, Oregon, USA
E-mail: beenish_z@hotmail.com

Citation: Zia B, et al. Synthetic Data for AI-Driven Detection of Laryngeal Cancer. J Glob Res Comput Sci. 2025;16:002

ABSTRACT

This paper explores the use of Artificial Intelligence (AI) for detecting laryngeal cancer through synthetically generated voice data, an area with limited research. While previous studies have shown promise using real voice data, they often rely on proprietary datasets, with no publicly available models for replication. The authors address this gap by developing and open sourcing an AI model trained on synthetic voice data to distinguish between normal and diseased voices. The paper compares the performance of the model using synthetic and real voice data, discusses the methods used for data preparation and model training, and presents the resulting open-source model for public use. This work contributes to the growing body of research on AI and voice data, offering a valuable resource for further exploration in medical diagnostics.

Keywords: Artificial intelligence; Diseased; Data; Training; Models

INTRODUCTION

Artificial Intelligence (AI) has significantly impacted healthcare, enhancing diagnostics, treatment, and operational efficiency by handling large volumes of data, such as medical imaging and text data. However, limited research has been conducted on using AI for voice data, particularly use of synthetically generated voice data in detecting medical conditions such as laryngeal cancer. Previous studies on real voice data have shown promising results, but models have often been trained on proprietary or limited datasets such as the IEEE FEMH Voice Data Challenge, with no publicly available resources for replication [1-3].

While the Bridge2AI project aims to address similar issues, there is still a lack of publicly accessible models or datasets. This paper fills these gaps by presenting an AI model trained on synthetic voice data to detect laryngeal cancer, comparing results from synthetic and real data, and offering the open-source model for public use.

MATERIALS AND METHODS

To build an AI model that takes in audio files and classifies them as normal or diseased the following steps were taken:

1. Data Preparation:
    - Data Collection
    - Data Preprocessing
2. AI Model development:
    - Selection
    - Training
    - Evaluation

Data preparation

Waveform Audio File Format (WAV) files are commonly used for AI models due to their high audio quality, simplicity in processing, and compatibility with machine learning frameworks. Unlike compressed formats like MP3, WAV files are uncompressed, ensuring detailed and accurate audio data. This is why the authors chose WAV files as the input format for their AI models.

For data collection, the Rainbow Passage, frequently used by otolaryngologists in the U.S. for evaluation of voice disorders, served as a consistent test across both synthetic and real data samples. Two data sets were generated:

Real data

- **Normal data:** Family and friends were recruited to record the Rainbow Passage.
- **Diseased data:** Due to a lack of access to individuals with laryngeal cancer and the absence of open audio data, the authors simulated cancer data from normal data to test their model.

Synthetic data

- **Normal Data:** Google Cloud Text-to-Speech API6 generated synthetic speech with various voices and accents.
- **Diseased Data:** A python-based script was created to simulate the effects of laryngeal cancer on voice by modulating waveforms, shifting formants, pitch shifting, and adding noise to simulate voice degradation.

This script used pydub for audio manipulation, librosa for pitch and formant shifting, and ffmpeg for file operations. The script was open sourced on GitHub4 and applied to the synthetic dataset to generate diseased data.

Data preprocessing

The authors set a sample rate of 16,000 Hz, standardizing audio length for effective machine learning analysis. A variable called Max_length defined the length of recordings, with 1 minute and 45 seconds being sufficient for most cases.

AI Model development

To build an AI model that processes WAV audio files to classify voice data into normal vs diseased, the following steps were followed:

- Loading of audio .wav files
- Feature extraction
- Preprocess the features (e.g. padding or truncating to a consistent length)
- Create a model to classify the data in 'norma' or 'diseased'
- Training the model using the preprocessed data

For the loading of the audio*.wav files the authors used librosa, a python package for audio analysis. The raw waveform data was used as input to the model that was developed using Tensor flow framework.

Model architecture, training and evaluation

The authors used model a 1D Convolutional Neural Network (CNN), which is well-suited for analyzing time-series data like audio waveforms.

Training

- The model is compiled using the Adam (short for Adaptive Moment Estimation) optimizer and binary cross-entropy loss function since it is a binary classification problem.
- The model is trained for 10 epochs with a batch size of 32. This can be changed to reduce the memory footprint of the trained model.

Model Evaluation: After training, the model was evaluated on the test set, and accuracy was printed.

RESULTS

Below are the results that were gathered after training the model on various datasets and then evaluated using a new dataset for accuracy (Tables 1 and 2).

Table 1. Summary table for the results achieved by training and testing the deep learning-based model on real or synthetic data.

Data Type	Training samples (normal, disease)	Test samples (normal, disease)	Accuracy
Synthetic	180 (90,90)	24 (12,12)	91.60%
Synthetic	910 (555,355)	24 (12,12)	87.50%
Real	27 (24,3)	5 (4,1)	80%
Real	26 (24,2)	6 (4,2)	83.33%

Table 2. Summary table for the results achieved by training and testing the deep learning-based model on mixed data types.

Data Type	Training Samples (normal, disease)	Test Samples (normal, disease)	Accuracy
Real	27 (24,1)	5 (4,1)	57%
Synthetic	0	2 (1,1)	
Real	27 (24,3)	5 (4,1)	71.40%
Synthetic	15 (24,3)	2 (1,1)	
Real	27 (24,3)	5 (4,1)	80%
Synthetic	15 (11,4)	3 (2,1)	
Real Modulated	2 (0,2)	2 (0,2)	
Pretend Disease	0	1 (0,1)	

Real=Actual recorded voices from individuals

Synthetic=Voice samples generated using an AI based text to speech Application Programming Interface (API)
Real Modulated=Voice modulation script run on an actual recorded voice data, mimicking changes in voice due to cancer

Pretend disease=Actual voice recorded from individuals pretending to have hoarse voice. This was added to check if you can trick the model.

DISCUSSION

An AI model's performance heavily depends on the quality of data it is trained on. However, gaining access to health records and recorded voices from patients is a significant challenge, especially in the context of developing AI models. Initiatives like the Bridge2AI consortium are working on building human voice databases, but publicly accessible versions are still unavailable. This study seeks to address this gap by exploring the potential of training an AI model using synthetic data. While generating "normal" synthetic data was relatively straightforward, generating synthetic "diseased" data proved to be much more complex. Laryngeal cancer, which can significantly affect the voice, manifests in various ways, including:

- Hoarseness due to changes in vocal cord structure
- Breathy voice caused by incomplete vocal cord contact during phonation
- Pitch shifts, either lower or irregular
- Tremor or strained voice caused by vocal fold tension

These known variations served as the basis for creating a script that generated synthetic diseased data by modulating synthetic normal data. While deep learning models like tacotron or wave net could provide a more

accurate simulation of laryngeal cancer, simpler methods like pitch shifting, formant alteration, and noise addition were used to approximate the effect.

Once the data was generated, pre-processing was the next crucial step. An audio file is essentially a series of samples representing a sound wave over time, and the sample rate dictates how often these samples are taken. The authors used a sample rate of 16,000 Hz, a common rate for human speech. For machine learning models, it's essential that the input data has consistent size, especially when processed in batches. By ensuring fixed-length audio files, the authors achieved uniform input size, which facilitated efficient model training.

Feature extraction

Feature extraction is a critical part of AI model development. For feature extraction, techniques such as spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and deep learning models were considered. MFCCs are widely used in speech processing, as they capture characteristics of an audio signal related to human perception. These coefficients include:

- Envelope of the spectral content, mapping frequencies to the Mel scale.
- Short-term power spectrum, providing a spectrogram of the audio.
- Logarithmic compression, compressing amplitude values to align with human perception of loudness.
- Mel filter bank, capturing energy in frequency bands relevant to human hearing.
- Discrete Cosine Transform, reducing dimensionality.

MFCCs effectively capture the timbral characteristics (e.g., pitch, tone) of an audio signal, which is why they are commonly used in speech recognition.

Advantages of using MFCCs

- They provide a compact and effective representation of an audio signal.
- They are well-aligned with human auditory perception, identifying patterns in sound rather than raw waveforms.
- The extracted features are well-understood and directly utilized in AI model development.

Disadvantages of MFCCs:

- They do not capture information outside the human ear's perception, potentially missing spectral features that could be useful for training.
- The model accuracy for synthetic data was not as high when using MFCCs, suggesting the need for more advanced techniques.

Deep learning with raw audio: An alternative approach is to input raw audio waveforms directly into deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). These models can automatically learn meaningful features from the raw data without the need for manual feature extraction.

Advantages of deep learning with raw audio:

- Deep learning models like CNNs can learn complex patterns in the audio waveform.
- These models can capture spectral information outside the range of human hearing, which may provide additional insights for classification.
- The model showed better accuracy for synthetic data when using raw audio.

Disadvantages

The model is less explainable, as the developer cannot easily see what features are being extracted by the deep learning techniques, making it harder to explain the model's decisions.

After experimenting with both MFCC and deep learning features, the authors chose deep learning for feature extraction, as it showed higher accuracy and was better aligned with the goal of discovering patterns in the audio data that humans might not perceive.

As mentioned before in this paper, the existing models for laryngeal disease classification have utilized proprietary models trained on closed databases. There were a few common themes across these works [1,2,3]:

- The models used in these studies were closed source, making replication impossible.
- The datasets, whether normal or diseased, were proprietary or available only for limited periods.
- No studies had published results on using synthetic data to train and test AI models for laryngeal disease classification.
- Little publication exists on using frequencies outside the human audible range for disease detection.
- In this study, the authors aimed to address these gaps by open sourcing their model on GitHub [4], along with detailed results from experiments conducted on synthetic and real data.

Experiments

1. **Synthetic data only**: (a) Smaller Training Data: In the first experiment, the model was trained on 180 samples (90 normal, 90 diseased) and tested on 24 samples (12 normal, 12 diseased). The model achieved an accuracy of 91.6%, detecting 22 out of 24 correctly.
   (b). Larger Training Data: In the second experiment, the training data increased to 910 samples (555 normal, 355 diseased). The model was tested on 24 samples and achieved an accuracy of 87.5%, detecting 21 out of 24 correctly. Interestingly, increasing the training data slightly reduced the accuracy, possibly because the difference between synthetic normal and diseased data was too subtle for the model to distinguish effectively.
2. **Real data only:** The authors were able to collect real normal data from volunteers and some diseased data using electro larynx devices. The model was trained on 27 samples (24 normal, 3 diseased) and tested on 5 samples (4 normal, 1 diseased), achieving 80% accuracy. The model failed to detect diseased voices correctly, likely due to the small diseased dataset. The results from real data and synthetic data only are captured in Table 1.
3. **Real data+Pretend disease:** The model was trained on 26 samples (24 normal, 2 diseased) and tested on 6 samples (4 normal, 1 real diseased, 1 pretend diseased). The accuracy was 83.33%, but the real diseased voice was misclassified. The model did not mistake the hoarse voice for diseased, indicating it could differentiate between normal and diseased voices more effectively.
4. **Mixed Data (Synthetic, real and modulated):** The authors conducted several experiments combining real, synthetic, and modulated data:

- **Training Real, Testing Real and Synthetic:** The model was trained on 27 real samples (24 normal, 3 diseased) and tested on 7 samples (5 real, 2 synthetic). The accuracy was 57%, with the model failing to detect diseased voices.

- **Training and Testing on Real and Synthetic Data:** The model was trained on 42 samples (35 normal, 7 diseased) and tested on 7 samples (5 real, 2 synthetic). The accuracy improved to 71.4%, with 5 out of 7 samples correctly classified.
- **Training and Testing on Real, Synthetic, and Modulated Data:** In this experiment, the model was trained on 44 samples (35 normal, 9 diseased) and tested on 10 samples (6 normal, 4 diseased). The model achieved an accuracy of 80%, with some misclassifications of modulated real data. The results from mixed data type training and testing are captured in Table 2.

The experiments showed that a model trained purely on synthetic data achieved the highest accuracy (>90%). However, such a model may not perform well on real clinical data, leading to false positives and negatives. A model trained on a mix of real and synthetic data achieved acceptable results, although the modulation of real voices to mimic diseased voices did not always work well. In conclusion, while synthetic data offers a viable path for AI training, mixing real and synthetic data provides a more robust model, but further work is needed to improve the simulation of real diseased voices.

The authors use of synthetic data to train and test the model, that has been open sourced [4], provides a new approach to use of large language model-based synthetic data generation use in cancer classification. It also highlights the need of open-source real data for creation of robust models. The authors believe that with better real cancer voice data, their model might be able to provide more impactful observations and hence are open sourcing their model as well as modulating script, so community can experiment with real and synthetic data and share results that eventually could be used to build a federated model for cancer detection or could be the basis for a foundation model that can be used for early detection of neurological diseases like Parkinson's Disease, mental health and others, using voice data.

CONCLUSION

This study demonstrates the potential of using synthetic voice data to train AI models for detecting laryngeal cancer. By generating and modulating synthetic speech, the authors addressed the scarcity of real patient data and achieved high accuracy with their deep learning model. While synthetic data alone performed well, combining it with real and modulated samples produced more balanced results. The open-source release of the model and modulation tools invites further research and collaboration. This work lays the groundwork for future voice-based diagnostics, potentially aiding early detection of laryngeal cancer and other conditions like Parkinson's disease and mental health disorders.

AUTHOR CONTRIBUTIONS

Conceptualization: BZ, FT; Data curation: BZ; Formal analysis: BZ; Methodology: BZ, FT; Investigation: BZ, FT; Visualization: BZ, FT; Funding acquisition: None; Project administration: BZ, FT; Resources: BZ, FT; Software: BZ; Supervision: BZ, FT; Validation: BZ, FT; Writing–original draft: BZ; Writing–review & editing: BZ, FT

COMPETING INTERESTS

Authors declare that they have no competing interests.

DATA AND MATERIALS AVAILABILITY

All data are available in the main text or the supplementary materials.

REFERENCES

1. Kim H, et al. Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy. J Clin Med. 2020;9:3415. [Crossref] [Google Scholar] [PubMed]
2. Kim HB, et al. Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis. Sci Rep. 2024:14:9297. [Crossref] [Google Scholar] [PubMed]
3. Ramalingam A, et al. IEEE FEMH Voice Data Challenge 2018. 2018 IEEE International Conference on Big Data (Big Data). 2018;5272-5276. [Crossref]
4. Beenish Z. Voice abnormality detection. 25 March, 2025.
5. Cancer Research. Electrolarynx after a laryngectomy. 20 March, 2025.
6. Google Cloud. Cloud Text-to-Speech. 20 January, 2025.