JOURNAL OF ACADEMIC RESEARCH FOR HUMANITIES
ISSN 2663-192x

ORCID of the Journal: https://orcid.org/0009-0000-0723-9485
DOI Number of the Paper: https://zenodo.org/records/14558219
Edition Link: Journal of Academic Research for Humanities JARH, 4(*4*) Oct-Dec 2024
Link of the Paper: https://jar.bwo-researches.com/index.php/jarh/article/view/530
HJRS Link: Journal of Academic Research for Humanities JARH (HEC-Recognized for 2023-2024)

EVALUATING RNN VARIANTS FOR DYSPHONIA CLASSIFICATION USING THE UNCOMMON VOICE DATASET: A COMPARATIVE ANALYSIS

Corresponding & Author 1:	**Irum Sindhu,** CS Lecturer, Sukkur IBA University, Sukkur, Sindh, Pakistan, Email: irum.sindhu@ibs-suk.edu.pk
Author 2:	Faryal Shamsi, CS Lecturer, Sukkur IBA University, Sukkur, Sindh, Pakistan, Email: faryal.shamsi@iba-suk.edu.pk
Author 3:	Muhammad Shamrie Sanin, Senior Lecturer, University Malaysia Saba, Malaysia. Email: Shamrie@ums.edu.my

Paper Information

Citation of the paper:
(Jarh) Sindhu, I., Shamsi, F., & Sanin, M, S., (2024). Evaluating RNN Variants for Dysphonia Classification Using the Uncommon Voice Dataset: A Comparative Analysis. In *Journal of Academic Research for Humanities*, 4(*4*), 113-122.

QR Code for the Paper:

Subject Areas for JARH:
1 Humanities
2 Business Administration

Timeline of the Paper at JARH:
Received on: 26-11-2024.
Reviews Completed on: 22-12-2024.
Accepted on: 26-12-2024.
Online on: 26-12-2024.

Recognised for BWO-R:
HJRS HEC Journal Recognition System

Published by BWO Researches INTL.:
BWO RESEARCHES (PVT) LTD INTERNATIONAL

DOI Image of the paper:
DOI 10.5281/zenodo.14558219

Abstract

Dysphonia, a voice disorder caused by vocal cord dysfunction, significantly affects individuals' communication abilities. Early and accurate detection of dysphonia is crucial for timely intervention and effective treatment. Leveraging advancements in deep learning, this study employs RNN - Recurrent Neural Networks to enhance detection accuracy. However, most studies in this domain rely on common datasets, leading to limited generalizability of models to diverse populations. In this study, we explore the use of various RNN variants, including traditional RNN - Recurrent Neural Network, LSTM - Long Short-Term Memory and GRU - Gated Recurrent Neural Network, to detect dysphonia in an uncommon voice dataset. Existing works focus primarily on conventional datasets and simpler classifiers, leaving room for improvement in accuracy and robustness. Our methodology leverages feature extraction techniques to preprocess the dataset, followed by training RNN variants to evaluate their performance in classifying dysphonic and non-dysphonic voices. Each RNN variant was trained and evaluated on the preprocessed dataset, divided into an 80:20 ratio for training and testing. The results revealed differences in model performance, with the standard RNN achieving an accuracy of 76%, while the LSTM and GRU models outperformed it, achieving accuracy of 94% and 93%, respectively. The experimental results demonstrate the effectiveness of advanced RNN models in handling diverse and challenging datasets, offering insights into their comparative performance for dysphonia detection and advancing research in voice disorder diagnosis.

Keywords: Dysphonia, Disorder, RNN, LSTM, GRU.

Introduction

Human voices play a crucial role in daily communication, allowing us to convey ideas, build connections, and express emotions. However, Dysphonia is a condition commonly referred to as hoarseness. It can significantly impair this essential function. Dysphonia includes a range of voice disorders marked by abnormal vocal qualities, such as hoarseness, breathiness, weakness, or complete loss of voice. Symptoms may develop suddenly or gradually and can include voice breaks, pitch instability, or even pain during speech. A distinct form of Dysphonia, Spasmodic Dysphonia (SD), is a neurological disorder that specifically affects the larynx muscles, causing involuntary spasms. These spasms lead to voice disruptions, producing a strained, strangled, or breathy quality. Understanding the unique causes and characteristics of SD is essential for devising effective treatments and improving the quality of life for affected individuals. Dysphonia, if not treated early, may cause further complications, leading to significantly hindering an individual's ability to communicate effectively, impacting their quality of life. Fortunately, researchers are actively working to improve speech recognition technology for individuals with voice disorders (Shamsi et al., 2019). Several research have been published to highlight the common issues in this field like the scarcity of datasets. A key tool in this effort is the application of advanced machine learning methods, especially deep learning models, to analyze vocal patterns and detect subtle abnormalities linked to Dysphonia in the "UncommonVoice" dataset. The "UncommonVoice" dataset contains a total of three thousand, six hundred and ninety-three (3693) instances of voice recordings. This provides a great source for developing and testing machine-learning models for various speech-related tasks (Moore et al., 2020). The dataset consisted of crowd-sourced speech recordings from 57 individuals with voice disorders, primarily focusing on Spasmodic Dysphonia (SD). This dataset is Developed in collaboration with Arizona State University's Center for Cognitive Ubiquitous Computing, to fill a gap by providing huge speech data specifically covering dysphonia. Participants were recruited with the support of the National Spasmodic Dysphonia Association. They contributed through web-based recordings, covering tasks ranging from sustained vowels to intelligibility assessments and image descriptions. Despite its potential, the dataset named "UncommonVoice" is not yet fully utilized for ongoing research on dysphonia detection, presenting an opportunity for exploring its full capabilities and contributing to the field's understanding of voice disorders. According to (Shamsi & Sindhu, 2024) basic RNNs, LSTMs, and GRUs are state-of-the-art types of deep neural networks designed to capture temporal dependencies in sequential data. This capability makes them highly appropriate for speech-processing tasks, where the sequence and timing of information are essential. However, traditional RNNs often struggle with capturing long-term relationships due to the peculiar concern of vanishing gradient. LSTMs overcome this challenge by incorporating memory cells and gating mechanisms that preserve information over longer sequences, while GRUs offer a simplified version of the LSTM with fewer parameters but comparable performance. Notably, to the best of our knowledge, no comprehensive comparative study of these RNN variants has been conducted specifically for speech sound disorders, underscoring the importance of the current research. This study aims to fill this gap by conducting a comparative analysis of standard RNNs, LSTMs, and GRUs for the task of dysphonia classification using the dataset named "UncommonVoice". By systematically evaluating the performance of these variants, we seek to identify the most effective model for this specific application and provide insights into the relative strengths and limitations of each variant. Notably this research fills a gap in

the comparative analyses of RNN variants for speech sound disorders, underscoring the novelty and importance of our work. Our results contribute to the growing body of research on speech disorder identification, offering potential improvements in voice-based diagnostic tools and assistive technologies.

Literature Review

Recurrent Neural Network

RNNs are a type of neural network designed to handle sequential data, where the order of information matters. They achieve this by incorporating a hidden state that carries information across processing steps. The basic RNN update equation for the hidden state is:

$$q_t = f\,(U_q * q_{t-1} + u_i * v_t + b_q) \quad (1)$$

Where:

f: Activation Function (*e.g., tanh, sigmoid*)

U_q: Weight matrix for the hidden state

U_i: Weight matrix for the input

b_q: Bias vector

q_t: hidden state at time step

This equation demonstrates how the hidden state q_t at time step t is influenced by the previous hidden state q_{t-1}, the current input u_i, and a bias term b_q. By applying an activation function, the network introduces non-linearities, enabling it to capture intricate patterns in the data. However, RNNs encounter challenges in preserving information over long sequences due to the issue of vanishing gradients, where gradients diminish as they propagate backwards through the network, potentially causing earlier time steps to have less impact. In a comparative study on voice pathology detection (Syed et al., 2021), CNN and RNN models were evaluated using the SVD dataset, demonstrating CNN's slightly higher accuracy of 87.11% compared to RNN's 86.52%. The study employed a complex architecture featuring 27 layers, combining convolutional and recurrent neural networks for feature extraction and analysis highlighting the need for further exploration of comparative analyses with other neural network variants. Another study (Ksibi et al., 2023) presents a deep learning approach for accurate detection of speech pathology, concentrating on single-vowel analysis (e.g., /a/) and omitting analysis of phrases and other vowels. The research introduces a novel CNN-RNN architecture tailored for voice pathology detection, achieving notable performance with an accuracy of 88.84% and an F1 score of 87.39%. However, different variants like GRU and LSTM were still left unexplored.

LSTM

LSTM, one of the variants of RNN, tackles the vanishing gradient problem by employing a sophisticated cell structure featuring gates (Graves et al.,2014). These gates regulate the flow of information inside the cell, enabling it to retain important information over extended sequences. LSTMs have similar components to RNNs, but their hidden state is replaced by a cell state and a hidden state is derived from the cell state. Additionally, LSTMs introduce three gates:

Forget Gate: Determines which information from the previous cell state to discard.

***Input Gate*:** Chooses which data from the current input should be retained in the cell state.

***Output Gate*:** Decides which data from the current cell state should be included in the hidden state output.

The LSTM update equations involve several calculations for each gate and the cell state:

$$f_t = \sigma(w_f * q_{t-1} + v_f * i_t + b_f\ ) \quad (2)$$

$$i_t = \sigma(w_i * q_{t-1}\ + v_i * i_t + b_i \quad (3)$$

$$\sim c_t = tanh(w_c * q_{t-1}\ + v_c * i_t + b_c \quad (4)$$

$$c_t = f_t + c_{t-1} + i_t * \sim c_t \quad (5)$$

$$o_t = \sigma(w_o * q_{t-1}\ + v_o * i_t + b_o \quad (6)$$

$$h_t = o_t + tanh(c_t) \quad (7)$$

These equations show how the gates regulate information flow. The forget gate f_t in eq (1) determines which information to remove from the cell state, the input gate in eq(3) chooses new information to store, and the output gate dictates what the network retains at the current time step. Various studies have used this LSTM with different feature sets and for various

pathologies. One study presents a deep learning approach using an LSTM autoencoder with multi-task learning to detect pathological voice disorders from continuous speech signals (Sztahó et al., 2021). It achieves high accuracies of 85% for Parkinson's disease, 86% for dysphonia, and 90% for depression across evaluation datasets. Another study compares SVM, BiLSTM, and CNN algorithms for detecting spasmodic dysphonia using MFCCs from the Saarbrucken Voice Database. BiLSTM and CNN achieved accuracies of 96.20%, outperforming SVM (96.15%), showing promise for automated detection of this voice disorder (Merzougui et al., 2024). Research has also been done to classify various dysphonia categories. This study (da Silva et al., 2024) categorizes vocal pathologies into functional, organic, and organofunctional types using the Saarbruecken Voice Database. It utilizes spectrogram-based classification with a Convolutional Neural Network (CNN). Results indicate that the CNN achieved 75.4% accuracy for organic dysphonia, 67.5% for functional dysphonia, and 52.9% for multi-label classification. As summarized by the author in her study (Sindhu et al., 2024) of systematic literature review, most of the researchers have used CNN for the classification of voice pathology.

Gated Recurrent Unit

According to the literature (Shih et al., 2022) GRUs reduce the complexities of LSTM architecture by incorporating single gates, removing a couple of gates i.e. Input Gate and the Forget Gate. The effective computation exploited by GRU architecture for determining the values of the gates is mathematically demonstrated in the following equations:

$$m_t = \sigma(w_z * q_{t-1} + v_m * i_t) + b_m \quad (8)$$

$$r_t = \sigma(w_r * q_{t-1} + v_r * i_t) + b_r \quad (9)$$

$$qt' = \tanh(w_h * ([r_t * q_{t-1} + i_t) + b_q \quad (10)$$

$$q_t = (1 - m_t) * q_{t-1} + m_t * qt' \quad (11)$$

Where m_t in eq(8) is the update gate that determines how much of the previous hidden state q_{t-1} to retain and how much to update with new information from the current input whereas w_z is the weight matrix for the previous hidden state, v_m is the weight matrix for input, it is the reset gate and qt' is the candidate hidden state. The candidate hidden state qt' is a combination of the reset gated hidden state $r_t * q_{t-1}$ and the transformed input. Eq (11) shows the final hidden state q_t, which is a weighted sum of the previous hidden state (q_{t-1}) and the candidate hidden state qt' controlled by the update gate m_t. These modifications enable GRUs to capture dependencies over longer sequences with fewer parameters, making them computationally efficient. Shewalkar et al. introduce a combined CNN-GRU model integrating convolutional neural networks and gated recurrent units for dysarthria detection (Shewalkar et al., 2019). Experimental findings indicate that the proposed CNN-GRU model achieves a leading accuracy of 98.38%, surpassing other models in the field. Apart from this GRU has been widely used in the task of speech recognition. The study (Shewalkar et al., 2018) & (Shewalkar et al., 2019) focuses on single-vowel analysis but does not evaluate the efficacy of GRUs or LSTMs, leaving a research gap. Findings indicate that LSTM achieves the lowest word error rates, although GRU optimization exhibits faster convergence while maintaining competitive word error rates like LSTM. Another study compares GRU and LSTM models for large vocabulary continuous speech recognition using TED talks (Khandelwal et al., 2016). The author concludes that GRU, simpler than LSTM, consistently outperforms LSTM across all network depths in speech recognition tasks. Apart from these deep learning models, spasmodic dysphonia is also classified using machine learning algorithms. The authors used three widely employed classifiers: k-nearest neighbours (KNN), Support Vector Machine (SVM), and Decision Tree (DT) on Saarbruecken Voice Database (SVD) (Hadjaidji et al., 2021). The Decision Tree algorithm achieved the highest classification accuracy, approximately 86.66%. Another study (Rivera et al., 2023) compared six machine learning algorithms for the automatic identification of dysphonia, with

KNN showing the best accuracy among all of them (87% - 92%).

Another class of models, RNNs, is particularly suited for sequential data like voice recordings in the realm of voice pathology detection, particularly in dysphonia, deep learning models such as RNN, LSTM, and GRU have not been extensively studied despite their successful application in various domains such as speech recognition and natural language processing. This study aims to fill this gap by thoroughly investigating the performance of these core RNN variants. By focusing on their basic structures and avoiding unnecessary complexity, we aim to uncover their potential for accurately classifying dysphonia. This research fills a crucial gap by systematically evaluating these RNN variants, providing insights into their efficacy in a domain where they have been underutilized. This exploration not only contributes to the field of voice disorder diagnostics but also lays the groundwork for future enhancements in voice pathology detection and treatment strategies.

Methodology

The study utilized the dataset named "UncommonVoice", which contains crowd-sourced voice recordings from 57 speakers with various speech disorders, primarily focusing on Spasmodic Dysphonia (SD). The "UncommonVoice" dataset underwent preprocessing steps to prepare it for model training. The preprocessing steps include: 1) Feature Extraction and 2) Standardization. Three recurrent neural network (RNN) variants—Standard RNN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)—were selected for comparative analysis.

1. *Standard RNN*

Configured with a single RNN layer containing 128 units, and 64 units at a hidden layer optimized for simplicity in capturing sequential dependencies.

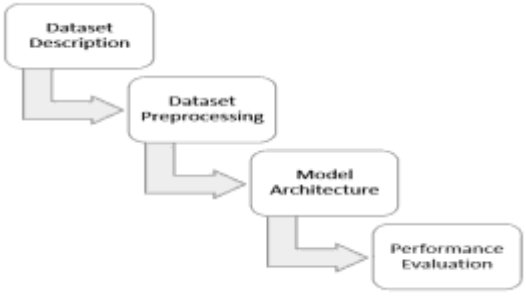
2.LSTM

The LSTM model incorporated 128 units, and 64 units at hidden layers leveraging its gating mechanisms to preserve long-term dependencies.

3.*GRU*

"The GRU model was designed with 128 units, and 64 units at hidden layers combining simplicity and efficiency for temporal modelling.

GRU Utilized a GRU layer known for its simplified architecture compared to LSTM, yet capable of capturing temporal dependencies effectively. Models were trained on an 80:20 split of the dataset, optimizing with categorical cross-entropy and Adam optimizer. Evaluation metrics included accuracy, precision, recall, and F1 score to gauge classification performance. The study maintained consistency in experimental settings across all RNN variants. Hyperparameters such as batch size, number of epochs, and model complexity were kept uniform to facilitate fair comparisons of their performance metrics. Figure 1 demonstrates the complete flow of methodology. Each step of the diagram is further explained in the sub-sections below.



1. Dataset Description

In this study, we have used the "UncommonVoice" dataset which is publicly available for research purposes. This data set contains speech recordings from 57 people, majorly those having Spasmodic Dysphonia (SD). Surveys were initially conducted with the participants, during which they were asked about the details of their voice condition and were asked to rate their voice quality.

Dataset Structure

i. Participant Information

a) *Demographics:* Includes age, gender, and other relevant details.

b) *Voice Condition:* Self-reported information about the presence and specifics of any voice disorders.

ii. Speech Recordings

a)*Non-Words:*

■ Sustained vowels (e.g., /a/, /i/, /u/).

■ Diadochokinetic (DDK) rate tasks to assess speech motor control.

b) *Read Speech:*

■ Sentences randomly selected from the TIMIT corpus.

■ Sentences required for the CAPE-V intelligibility assessment.

c) Spontaneous Speech:

■ Descriptions of images sourced from the MSCOCO dataset.

iii. Recording Details:

a) Participants recorded their speech using a web-based application, utilizing their recording equipment.

iv.Data Format

a) Audio Files: Stored in standard formats (e.g., WAV) with appropriate sampling rates to ensure quality and compatibility.

2. Dataset Preprocessing

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used features in speech and audio processing (Abdul et al., 2022), particularly in tasks like speech recognition, speaker identification, and music classification. They represent the short-term power spectrum of sound in a way that mimics human auditory perception. MFCCs are computed by applying a series of transformations, including Fourier transform, Mel filter bank, logarithmic compression, and discrete cosine transform (DCT), to capture the essential characteristics of audio signals. These features effectively model the frequency content of sound, making them robust and efficient for various sound classification applications. To achieve quality and to make the data compatible with the deep learning models, we performed several preprocessing steps including feature extraction, dataset balancing, label encoding and dataset splitting. The first step involved extracting features from each audio file. In this study, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted which capture the spectral features required for speech analysis. In this process, each audio file is initially segmented into frames. Afterwards, the power spectrum of each frame is represented by calculating its coefficients. In our study, we calculated the 13 coefficients, which are sufficient to represent each audio segment. After the feature extraction, we performed data balancing which was required to achieve uniformity in data dimensions. To achieve the constant frame size of 300, extracted MFCCs were truncated to maximum length or padded with zeros. This step was necessary because each audio file had different lengths, thereby facilitating seamless integration into the training pipeline. The reason is to select 300 frames for the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) to ensure an adequate representation of the audio signal while maintaining computational efficiency. Each frame typically spans 20-25 milliseconds, which allows us to capture important temporal variations in the acoustic features of the signal. The choice of 300 frames corresponds to a segment of approximately 7.5 seconds, which is long enough to encompass meaningful speech or audio patterns while remaining within a practical processing window. This duration provides sufficient information for downstream analysis, such as speech recognition or classification tasks, without introducing unnecessary computational complexity. Additionally, the fixed number of frames ensures consistency in feature representation, allowing for easier comparison across different samples and facilitating the training of machine learning models. Thus, 300 frames offer a balanced trade-off between the need for rich temporal features and the desire for manageable input data for model training. After performing the dataset balancing, next we structure the data into arrays. The arrays were composed of MFCC features and binary labels 0

and 1 (0 for absence and 1 for presence of dysphonia). These labels indicated the presence and absence of dysphonia, or we can say normal and abnormal speech. This labelling is done to satisfy the requirement of the deep learning models which are designed to perform binary classification tasks. To check the performance of the model, the dataset is divided into training and test sets. The 80:20 stratified technique is used, with 80 per cent dedicated to training sets and 20 per cent data reserved for testing purposes.

3. Model Architecture

We have proposed three model architectures representing the variants of RNN. The architecture details of each model and experimental settings are described below in this section. Experimental Settings and Environment All three model architectures were evaluated under the following experimental settings:

Hardware:

- Processor: Dell core i7 3m4 GHz 6700 cpu
- RAM: 8 GB memory

Software:

- TensorFlow 2.0 with the Keras API: TensorFlow 2.0, with the integrated Keras API, provides an intuitive and efficient framework for implementing and training machine learning models, particularly deep learning models.
- Librosa: Librosa is a Python library for audio and music analysis. It provides tools for performing tasks such as loading audio files, extracting features like Mel spectrograms, and performing transformations on audio signals. In the context of speech and audio processing, Librosa is often used for preprocessing and feature extraction, such as generating Mel-Frequency Cepstral Coefficients (MFCCs) from raw audio data.
- Numpy: Numpy is a core Python library for numerical computing, providing support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to manipulate these arrays.
- Matplotlib: Matplotlib is a widely used Python library for data visualization. It allows for the creation of static, interactive, and animated plots. In the context of machine learning and audio analysis, Matplotlib is often used to visualize training results, such as plotting loss curves, accuracy trends, and spectrograms, providing a clear visual representation of model performance and data characteristics.

Architecture Details

To achieve uniformity in the experiments, we designed the model with similar architectural components. Further, the models were designed with minimal layers to reduce model complexity. All three models namely: Simple RNN, Gated Recurrent Unit (GRU), and Long Short Term Memory (LSTM) models were configured with one model layer, one hidden layer and output layer. The purpose of these minimal layer models was to check how they perform without introducing any complexity. In the RNN model, the first Simple RNN layer with 128 units, a dense hidden layer with 64 units, and an output layer representing binary classification was used. In this study, we utilized a Recurrent Neural Network (RNN) with a configuration of 128 units in the first hidden layer and 64 units in the second hidden layer for speech recognition tasks on the UncommonVoice dataset. The UncommonVoice dataset, known for its diverse set of non-native speech data, presents unique challenges such as varying accents, noise, and speech patterns. The choice of 128 units in the first hidden layer allows the model to capture the complex temporal dependencies inherent in speech, particularly useful for dealing with the dataset's varied phonetic structures and non-native speech characteristics. The 64 units in the second hidden layer serve to refine these learned representations, reducing model complexity and enhancing generalization by preventing overfitting specific speaker characteristics. This configuration offers a balanced approach that ensures both sufficient capacity for accurate feature extraction and

efficient model training, critical for handling the diverse and challenging nature of the UncommonVoice dataset. In total, simple RNN has 26,562 trainable parameters. The GRU model featured a GRU layer with 128 units, followed by a dense hidden layer with 64 units, and an output layer for binary classification, resulting in 63,298 trainable parameters. The third model LSTM also used an LSTM layer with 128 units, a dense hidden layer and an output layer having 81,090 parameters. Initially, the epoch size was kept at 50 for all three models despite this setting, the training of the RNN stopped at 40 epochs, while the LSTM and GRU models halted at 24 epochs as shown in the training loss and validation loss graphs of each model. A batch size of 32 was selected because it speeds up the training process and results in memory efficiency. The Adam optimizer, evaluated with categorical cross entropy, was used as the loss function. To introduce non-linearity in the model and understand the complex patterns of shorter and longer sequences ReLU activation function was utilized, respectively. All models were introduced with L2 regularization with a parameter of 0.01 to prevent overfitting. Additionally, early stopping with a patience of 5 epochs was employed during training to halt the training process if validation loss did not improve, thereby preventing overfitting and ensuring the best-performing model parameters were retained. These standardized settings made it suitable to compare and identify the best model for the task of dysphonia classification. Table 1 summarizes the information regarding model architecture.

4. Performance Evaluation

To evaluate the performance of different variants of RNN in the task of dysphonia classification, we used multiple performance metrics. The first metric used is accuracy, which is the indicator of overall model prediction correctness. It measures the proportion of correctly classified instances out of the total predicted instances of the model. Precision was used to check the model's ability to correctly identify positive instances of dysphonia, excluding false positive instances. Recall on the other hand checks the model's capability that how many positive cases were identified by the model within the dataset. The F1 score, which is the combination of both precision and recall, provides a balanced evaluation of the model by identifying positive and negative dysphonia cases. Using these metrics in this study helps us to rigorously assess the performance of GRU, Simple RNN, and LSTM models, indicating their strengths and limitations in dysphonia classification.

Results and Discussion

This study attempted to evaluate the performance of 3 RNN variants. Those variants were Simple RNN, LSTM, and GRU, the analysis targeted the undertaking of dysphonia classification in the dataset named "UncommonVoice".

Table 1: Model Architecture

Model	Architecture	Parameters
GRU	GRU layer (128 units) → Dense hidden layer (64 units) → Output layer for binary classification.	63,298
RNN	Simple RNN layer (128 units) → Dense hidden layer (64 units) → Output layer.	26,562
LSTM	LSTM layer (128 units) → Dense hidden layer (64 units) → Output layer.	81,090

Fig. 2: Training and validation loss of each model

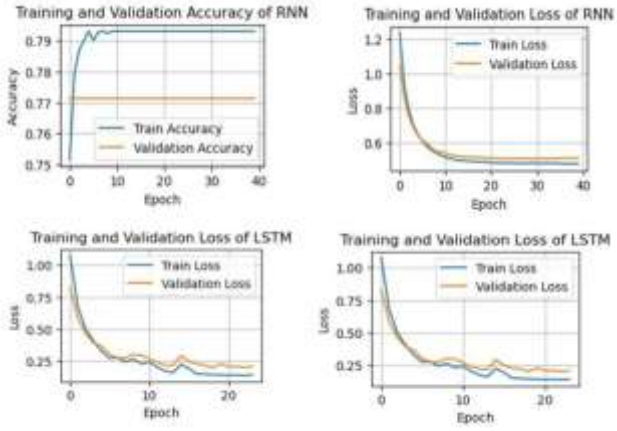


Table 2: Results

Model	Accuracy	F1 Score	Precision	Recall
RNN	77%	0.5949	0.7713	0.6717
LSTM	94%	0.9396	0.9425	0.9390
GRU	93%	0.9390	0.9391	0.9372

An accuracy of 77% was achieved by the Simple

RNN variant. Its Precision and Recall were found to be 77.13% and 67.17% respectively. Moreover, the F1 Score was calculated as 59.4. Although the Simple RNN has a straightforward architecture, it achieved moderate performance as compared to other models in this task. This outcome can be attributed to the architecture's lower complexity, having few parameters and simpler computations. Despite its easy implementation and faster training time, RNNs suffer from the vanishing gradient problem due to which the model could not capture long-term dependencies in this specific speech disorder. The LSTM model demonstrated significantly higher performance with an accuracy of 0.94, indicating that 94% of the model's predictions were accurate. The precision value was 0.9408 and the recall was 0.9390, indicating that the LSTM correctly identified dysphonia instances with high precision and sensitivity. The F1 score of 0.94075 further confirms the model's balanced performance, reflecting its ability to maintain high precision while effectively capturing all positive instances in the dataset. The LSTM's ability to capture long-term dependencies in sequential data proved advantageous in accurately identifying instances of dysphonia, showcasing its effectiveness in this classification task. However, as shown by the complexity of the architecture (more parameters), LSTM is known for its high computational cost, longer training times, and increased memory requirements. Similarly, the GRU model achieved an accuracy of 0.93, demonstrating its strong performance in dysphonia classification. Precision was 0.9390, indicating that when the model predicted dysphonia, it was correct 93.90% of the time. The 93.91% (almost 94%) recall value highlights the GRU model's strong capability to identify a significant proportion of positive instances in the dataset. With an F1 score of 0.9372, the GRU effectively balances precision and recall, offering a well-rounded evaluation of its performance. Its efficiency in training and adeptness at handling sequential data further bolstered its competitive performance, rivalling the LSTM. These findings emphasize the superior accuracy and robustness of LSTM and GRU models compared to the Simple RNN in classifying dysphonia using voice recordings. The superior accuracy and robustness of LSTM and GRU models highlight their suitability for tasks demanding precise classification of voice disorders. LSTM's ability to capture and retain long-term dependencies in sequential data, combined with the GRU's training efficiency and effective handling of sequential information, underpin their exceptional performance in this domain. Future studies could focus on advancing model architecture or refining feature engineering techniques to enhance classification accuracy and improve generalization across diverse datasets and conditions.

Conclusion

This study examined three types of recurrent neural networks (RNNs)—Simple RNN, LSTM, and GRU—for classifying dysphonia using the "UncommonVoice" dataset. The findings revealed that LSTM and GRU significantly outperformed Simple RNN in all evaluation metrics, including accuracy, precision, recall, and F1 score. Simple RNNs struggle with learning long-term dependencies. This limitation makes it challenging for Simple RNNs to capture temporal relationships in sequential data like speech. The LSTM model achieved 93% accuracy, while the GRU performed slightly better with 94% accuracy. These scores seem sufficient for this domain, as they demonstrate a significant ability to differentiate between normal and disordered speech, especially given the challenges associated with variability in voice data. However, while these results are promising, it is important to note that the real-world applicability depends on additional factors, such as generalizability across different datasets, robustness to noise, and integration with clinical workflows. This highlights the need to choose neural network architectures that are

well-suited to the challenges of sequential data, ensuring dependable and practical results.

Recommendations

Future research could focus on unravelling the internal mechanisms of GRU variants to better understand their performance in dysphonia classification. One area of interest is the examination of activation sequences, such as:

1. Candidate → Reset → Update → Forget → Activation → Output

2. Candidate → Update → Forget → Activation → Reset → Output → Activation

Analyzing these sequences may provide insights into how specific gate configurations impact the learning process and classification accuracy. Comparative studies on GRU models with modified gate arrangements or alternative activation functions could further optimize these architectures, making them more efficient and effective for voice disorder detection. Additionally, the research scope can be expanded to apply these models to diverse datasets that cover a variety of voice disorders like the Saarbruecken voice Database, or UA speech dataset that covers dysarthria. This would help to assess the generalization capabilities of these models. Moreover, developing methods to improve cross-condition and cross-population performance could enhance the reliability and applicability of dysphonia classification systems. These advancements could lead to the development of optimized RNN architectures specifically designed for voice disorder detection, paving the way for improved clinical and practical applications.

References

Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10, 122136-122158.

A. Graves, Long Short-Term Memory Networks with Residual Connections for Speech Recognition. 2014. Shamsi F, Sher MD, Shaikh S. Content-based automatic video genre identification. International Journal of Advanced Computer Science and Applications. 2019.

Dávid Sztahó, K. G., & Gábriel, T. M. (2021). 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 (pp. 135-141).

da Silva Moura, R., Maia, J. M., & Dajer, M. E. (2022, October). Detection and Classification of Categories of Dysphonia Using Convolutional Neural Network. In Latin American Conference on Biomedical Engineering (pp. 599-610). Cham: Springer Nature Switzerland.

Hadjaidji, E., Korba, M. C. A., & Khelil, K. (2021, September). Spasmodic dysphonia detection using machine learning classifiers. In 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI) (pp. 1-5). IEEE.

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

Moore, M., Papreja, P., Saxon, M., Berisha, V., & Panchanathan, S. (2020, October). Uncommonvoice: A Crowdsourced dataset of dysphonic speech. In Interspeech (pp. 2532-2536).

Merzougui, N., Korba, M. C. A., & Amara, F. (2024, January). Diagnosing Spasmodic Dysphonia with the Power of AI. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 1042-1046). IEEE.

RIVERA, M. A. B., GARCÍA, C. A. R., ROJAS, T. C. T., FLORES, P. M. Q., & LOAIZA, R. E. P. (2023). AUTOMATIC IDENTIFICATION OF DYSPHONIAS USING MACHINE LEARNING ALGORITHMS. Applied Computer Science, 19(4), 14-25.

Syed, Sidra Abid, et al. "Comparative analysis of CNN and RNN for voice pathology detection." BioMed Research International 2021.1 (2021): 6635964.

Schmidhuber, "Neural Networks for Compressing Sequences and Applications in Text and Speech Recognition," in Field Guide to Neurocomputing, Amsterdam, Netherlands: Elsevier, 2000, pp. 917–921.

Sindhu, I., & Sainin, M. S. (2024). Automatic Speech and Voice Disorder Detection using Deep Learning Systematic Literature Review. IEEE Access.

Shih, D. H., Liao, C. H., Wu, T. W., Xu, X. Y., & Shih, M. H. (2022, October). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. In Healthcare (Vol. 10, No. 10, p. 1956). MDPI.

Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. Journal of Artificial Intelligence and Soft Computing Research, 9(4), 235-245.

Shewalkar, A. N. (2018). Comparison of rnn, lstm and gru on speech recognition data.

Shamsi, F., & Sindhu, I. (2024). Condensing Video Content: Deep Learning Advancements and Challenges in Video Summarization Innovations. International" Journal of Academic Research for Humanities", 4(4), 113-124.

Shamsi, F., Sher, M. D., & Shaikh, S. (2019). Content-based automatic video genre identification. International Journal of Advanced Computer Science and Applications, 10(6).

S. Khandelwal, B. Lecouteux, and L. Besacier, Comparing GRU and LSTM for automatic speech recognition. Diss. LIG, 2016.