Journal of Intelligent Computing

DLINE JOURNALS

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2025: 16 (3)
https://doi.org/10.6025/jic/2025/16/3/114-124

# Evaluating RNN Variants for Dysphonia Classification using the Uncommon Voice Dataset: A Comparative Analysis

Irum Sindhu[1], Mohd Shamrie, Sanin[2]
[1]University Malaysia Sabah 88400 Kota Kinabalu, Sabah Malaysia
https://www.ums.edu.my/v5/

[2]University Malaysia Sabah 88400 Kota Kinabalu, Sabah Malaysia

**ABSTRACT**

*Dysphonia, a voice disorder characterized by abnormal vocal quality, significantly impacts communication abilities. Accurate and early detection is crucial for effective treatment and intervention. This study compares the efficacy of various Recurrent Neural Network (RNN) variants in classifying dysphonia using the Uncommon Voice dataset and provides an evaluation of standard RNN, Gated Recurrent Unit (GRUs) and Long Short-Term Memory (LSTM) models. Each variant was trained and tested on the preprocessed dataset, split into 80:20 ratio of training and testing sets. The finding shows variations in model performance, where the standard RNN achieved an accuracy of 76%, while the LSTM and GRU models demonstrated superior accuraci -es of 94% and 93%, respectively. These results underscore the potential of advanced RNN variants, partic ularly LSTM and GRU, for dysphonia detection and classification. The analysis offers preliminary information n about the relative advantages and disadvantages of each RNN variant, paving the way for future resarch in the broader domain of speech sound disorder identification.*

**Keywords**: Dysphonia, Voice Disorder, Deep Learning, RNN, LSTM, Gated Recurrent Unit (GRU), Speech Sound Disorder Detection

**Received:** 5 March 2025, Revised 2 June 2025, Accepted 14 June 2025



## 1. Introduction

Our voices are central to everyday communication, enabling us to share ideas, build relationships, and express ourselves. However, a condition known as Dysphonia, also commonly referred to as hoarseness, can disrupt this vital function. Dysphonia encompasses a spectrum of voice disorders characterized by abnormal vocal quality, ranging from hoarseness and breathiness to weakness and complete voice loss. These changes can be sudden or gradual and may include voice breaks, pitch variations, or even pain while speaking. One specific type of Dysphonia is Spasmodic Dysphonia (SD). Unlike other forms of Dysphonia, SD is a neurological disorder affecting the muscles in the larynx, or voice box, causing them to spasm involuntarily. These spasms manifest as breaks in the voice, a strained or strangled quality, and breathiness. Understanding the specific causes and characteristics of SD is crucial for developing effective treatment strategies. This condition significantly hinders an individual's ability to communicate effectively, impacting their quality of life. Early and accurate diagnosis is crucial, as delayed intervention can lead to further complications.

Fortunately, researchers are actively working to improve speech recognition technology for individuals with voice disorders. One valuable resource in this endeavor is the Uncommon Voice dataset. The Uncommon Voice dataset, con-taining 3,693 instances of voice recordings, offers a rich resource for develop-ing and testing machine learning models for various speech-related tasks [1]. It is composed of crowd-sourced speech recordings from 57 individuals with voice disorders, primarily focusing on Spasmodic Dysphonia (SD). Developed in collaboration with Arizona State University's Center for Cognitive Ubiquitous Computing, this dataset aims to fill a niche by providing extensive speech data specifically from individuals with dysphonia. Participants, recruited with the support of the National Spasmodic Dysphonia Association, contributed through web-based recordings, covering tasks ranging from sustained vowels to intelligibility assessments and image descriptions. Despite its potential, the Uncommon Voice dataset remains underutilized in dysphonia classification research, presenting an opportunity for exploring its full capabilities and contributing to the field's understanding of voice disorders. Standard $RNN_s$, $LSTM_s$, and $GRU_s$ are advanced classes of deep neural networks that have been designed for identifying temporal connections in sequential data. This makes them particularly adept at handling speech processing tasks, where the order and timing of information are crucial. Despite their usefulness, traditional RNNs frequently have trouble identifying long term dependencies because of the vanishing gradient issue. $LSTM_s$ address this limitation by introducing memory cells and gating mechanisms to preserve information over extended sequences, while $GRU_s$ simplify the LSTM architecture with fewer parameters and comparable performance. Importantly, to the best of our knowledge, there has not been a comprehensive comparative analysis of these RNN variants specifically for speech sound disorders, highlighting the significance of the study in this paper.

This study aims to fill this gap by conducting a comparative analysis of standard $RNN_s$, $LSTM_s$, and $GRU_s$ for the task of dysphonia classification using the Uncommon Voice dataset. 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.

## 2. Related Work

$RNN_s$ 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:

$$h_t = f(w_h * h_{t-1} + w_x * x_t + b) \tag{1}$$

where:
$f$ : Activation function(e.g., tanh, sigmoid)
$zx_h$ : Weight matrix for the hidden state
$w_x$ : Weight matrix for the input
$b$: Bias vector
$h_t$: hidden state at time step

This equation demonstrates how the hidden state at time step t is influenced by the previous hidden state , the current input, and a bias term. By applying an activation function, the network introduces non-line arities, 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 backward through the network, potentially causing earlier time steps to have less impact. In a comparative study on voice pathology detection [2], 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 [3] 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, one of the variants of RNN, tackles the vanishing gradient problem by employing a sophisticated cell structure featuring gates [4]. 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:

$$f_t = \sigma(w_f * h_{t-1} + U_f * x_t + b_f \tag{2}$$

$$i_t = \sigma(w_i * h_{t-1} + U_i * x_t + b_i \tag{3}$$

$$\sim c_t = tanh(w_c * h_{t-1} + U_c * x_t + b_c \tag{4}$$

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

$$o_t = \sigma(w_o * h_{t-1} + U_o * x_t + b_o \tag{6}$$

$$h_t = o_t + tanh(c_t) \tag{7}$$

These equations show how the gates regulate information flow. The forget gate $f_t$ determines which information to remove from the cell state, the input gate $i+t$ chooses new information to store, and the output gate $o_t$ dictates what the network retains at the current time step. Various studies have used this LSTM with different features set and for various pathologies. One study presents a deep learning approach using an LSTM auto encoder with multi-task learning to detect pathological voice disorders from continuous speech signals [ 5]. 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 [6]. Research has also been done to classify various dysphonia categories. This study [7] categorizes vocal pathologies into functional, organic, and organo functional 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 [8] of systematic literature review, the majority of the researchers have used CNN for classification of voice pathology. Gated Recurrent Units (GRU) simplify the LSTM architecture by combining the forget and input gates into a single update gate and eliminating the output gate [9]. The GRU equations are:

$$z_t = \sigma(w_z * h_{t-1} + U_z * x_t + b_z \qquad (8)$$

$$r_t = \sigma(w_r * h_{t-1} + U_r * x_t + b_r \qquad (9)$$

$$h'_t = tanh(w_h * ([r_t * h_{t-1} + x_t) + b_h \qquad (10)$$

$$h_t = (1 - (z_t) * h_{t-1} + z_t * h'_t \qquad (11)$$

Where $z_t$ is the update gate, $r_t$ is the reset gate, and $h'_t$ is the candidate hidden state. These modifications enable GRUs to capture dependencies over longer sequences with fewer parameters, making them computationally efficient. Author in his study introduces a combined CNN-GRU model integrating convolutional neural net works and gated recurrent units for dysarthria detection [10]. 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. Author assesses the performance of RNN, LSTM, and GRU models using a downscaled TED-LIUM speech dataset in his study [11] [12]. Findings indicate that LSTM achieves the lowest word error rates, although GRU optimization exhibits faster convergence while maintaining competitive word error rates similar to LSTM. Another study compares GRU and LSTM models for large vocabulary continuous speech recognition using TED talks [13]. 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. Authors used three widely employed classifiers—k-nearest neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT) on Saarbruecken Voice Database (SVD) [14]. The Decision Tree algorithm achieved the highest classification accura -cy, approximately 86.66%. Another study [15] compared six machine learning algorithms for automatic identification of dysphonia, with KNN showing the best accuracy among all of them (87%-92%).

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 ground work for future enhancements in voice pathology detection and treatment strategies

## 3. Methodology

The study utilized the Uncommon Voice dataset, which contains crowd-sourced voice recordings from 57 speakers with various speech disorders, primarily focusing on Spasmodic Dysphonia (SD). The Uncommon Voice dataset underwent preprocessing steps to prepare it for model training. This included. feature extraction and standardization. Three recurrent neural network (RNN) variants—Standard RNN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)—were selected for comparative analysis. Each model's architecture was configured as follows: Standard RNN Configured with a simple RNN layer. LSTM Structured with an LSTM layer to capture long-term dependencies in the voice data. 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. Hyper parameters such as batch size, number of epochs, and model complexity were kept uniform to facilitate fair comparisons of their performance metrics. Figure 1 demonstrate the complete flow of methodology. Each step of the diagram is further explained in the below sub sections.

### 3.1 Dataset Description

The Uncommon Voice dataset is a publicly available collection of speech recordings contributed by 57 speakers, primarily focusing on individuals with voice disorders such as Spasmodic Dysphonia (SD). Participants, both with and without voice disorders, completed surveys detailing their voice conditions and provided self-reported ratings of voice quality. The data collection comprised tasks involving sustained corner vowels, DDK rate measurements, reading sentences from the TIMIT corpus, performing CAPEV intelligibility assessments, describing images from the MSCOCO dataset, and additional non-word tasks to track changes in voice over time.

### 3.2 Dataset Preprocessing

Before feeding the data into deep learning models, rigorous preprocessing steps were essential to ensure the quality and compatibility of the data. Each audio file was initially processed to extract Mel-Frequency Cepstral Coefficients (MFCCs),which are effective in capturing spectral features essential for speech analysis. MFCC extraction involves segmenting audio signals into frames and computing coefficients that represent the power spectrum of each frame. This process computed 13 coefficients per frame, offering a comprehensive representation of each audio segment. To ensure uniformity in data dimensions for model compatibility, extracted MFCCs were standardized by either truncating to a maximum length or padding with zeros to achieve a consistent frame size which is 300 in our case. This step was crucial in handling variable-length audio inputs, thereby facilitating seamless integration into the training pipeline.

Following feature extraction and standardization, the dataset was structured in to arrays containing the extracted MFCC features, and comprising binary labels indicating the presence or absence of dysphonia. The labels were encoded to facilitate binary classification, aligning with the requirements of the deep learning framework. To assess the model's performance, we partitioned the dataset in to training and test sets using an 80-20 stratified split, ensuring 80% of the data was allocated for training and 20% for testing. This split was executed with a fixed random state of 42 to ensure consistency in the evaluation process. This approach provided a robust measure of the model's ability to generalize unseen data and detect dysphonia effectively. Overall, these preprocessing steps, including feature extraction, standardization, label encoding, and train-test splitting, were crucial in preparing the Uncommon Voice dataset for binary classification tasks focused on dysphonia detection.

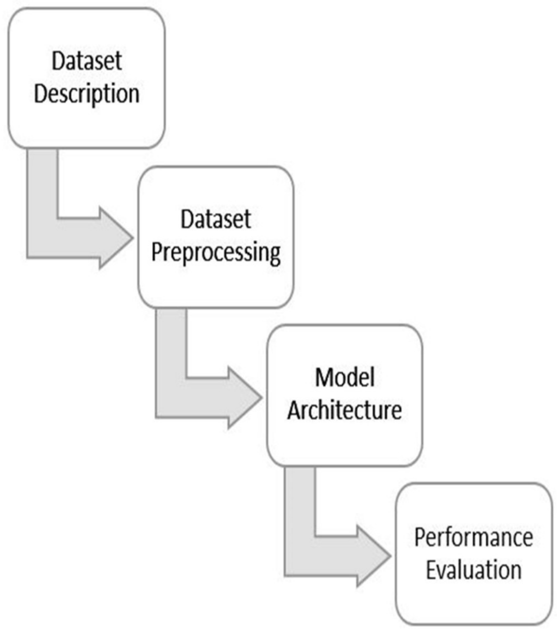


Figure 1. Methodology 3.1

### 3.3 Model Architecture

In this study, we evaluated three Recurrent Neural Network (RNN) variants for dysphonia classification using the Uncommon Voice dataset.

**Experimental Settings and Environment The experiments were conducted with the following settings:**

**Hardware:** The models were trained on an Dell core i7 3m4 GHz 6700 cpu with 8 GB memory which enabled efficient training and computation.

**Software:** The models were implemented using Tensor Flow 2.0 with the Keras API for building and training the neural networks. Librosa was used for preprocessing the audio data, and NumPy for handling data arrays. Results were visualized using Matplotlib.

Each model was configured with similar architectural components and minimal layers to ensure a consistent experimental setup. The models assessed in this study include a Gated Recurrent Unit (GRU) model, a Simple

RNN model, and a Long Short-Term Memory (LSTM) model, each configured with specific architectural details. The GRU model utilized a GRU layer with 128 units followed by a dense hidden layer comprising 64 units, culminating in an output layer for binary classification. This architecture resulted in a total of 63,298 trainable parameters. Similarly, the Simple RNN model consisted of a Simple RNN layer with 128 units followed by a dense hidden layer with 64 units and an output layer, totaling 26,562 trainable parameters. The LSTM model employed LSTM cells with 128 units followed by a dense hidden layer and an output layer, contributing to a total of 81,090 trainable parameters. Each model was trained for 50 epochs with a batch size of 32, optimized using the Adam optimizer, and trained using categorical cross-entropy as the loss function. ReLU activation function was also utilized in these models for its ability to introduce non-linearity and capture complex patterns within shorter and longer sequences, respectively. To mitigate over fitting, L2 regularization with a regularization parameter of 0.01 was applied to all model layers. 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 over fitting and ensuring the best performing model parameters were retained. However, each model was initially set to train for 50 epochs, 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. These standardized settings allowed for a direct comparison of their performance in dysphonia classification, aiming to identify the most effective model architecture for this specific application. Table 1 summarizes the information regarding model architectures.

| **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 |

Table 1. Model Architecture

### 3.4 Performance Evaluation

In evaluating the performance of various RNN variants for dysphonia classification using the Uncommon Voice dataset, this study employed multiple performance metrics to assess each model's efficacy. Accuracy served as a primary indicator of overall prediction correctness, measuring the proportion of correctly classif ied instances out of the total predictions made by the model. Precision was utilized to evaluate the model's ability to correctly identify positive instances of dysphonia, minimizing false positives. Recall measured the model's sensitivity in correctly capturing all positive instances within the dataset. Additionally, the F1 score, which combines precision and recall into a single metric, provided a balanced evaluation of the models' predictive capabilities across both positive and negative classes of dysphonia. Through the systematic application of these metrics, this study aimed to provide a comprehensive analysis of the GRU, Simple RNN, and LSTM models, offering insights into their respective strengths and limitations in the context of dysphonia classification.

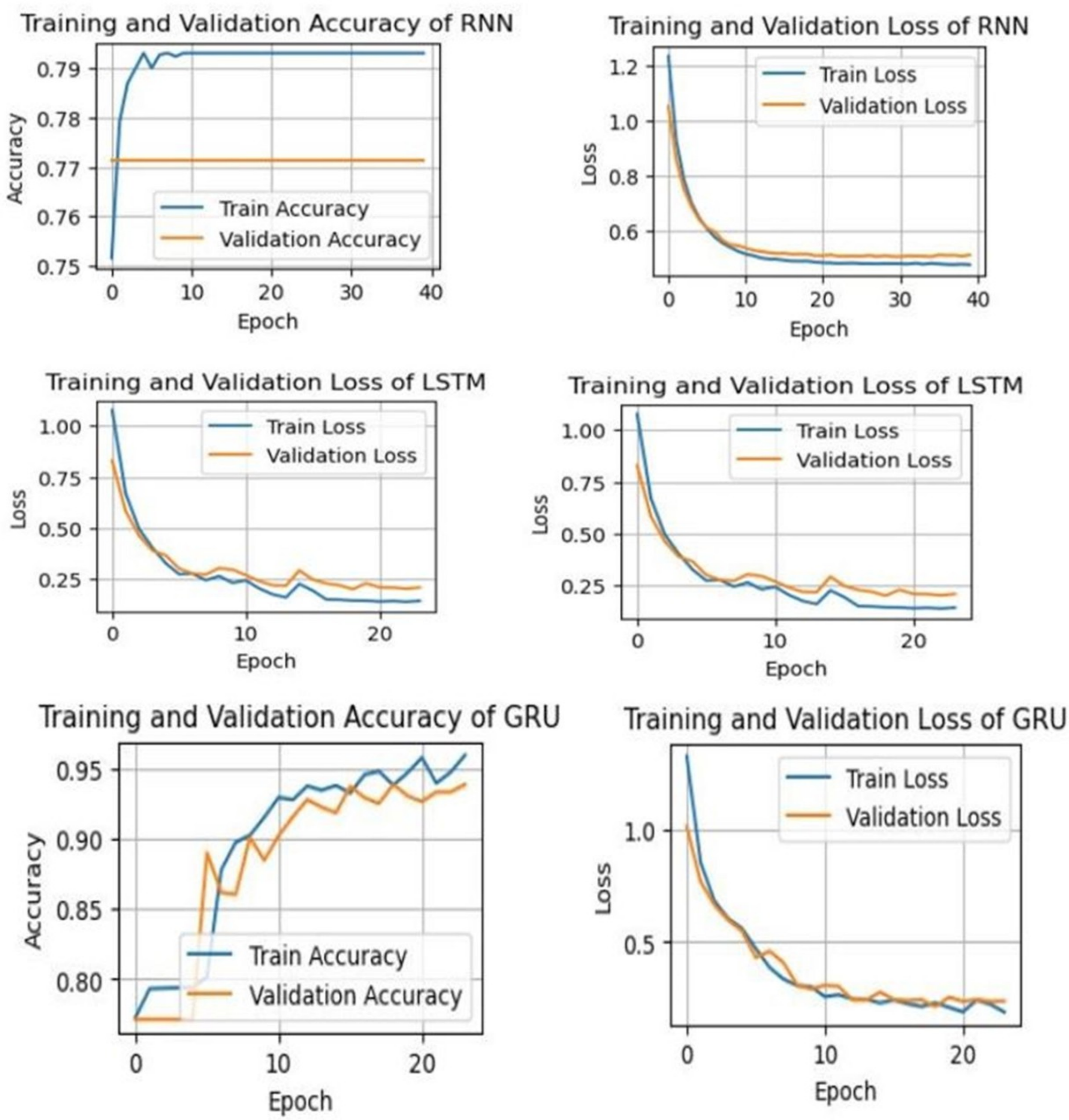


Figure 2. Training and validation Accuracy and loss of each model

## 4. Results and Discussion

In this study, we evaluated the performance of three recurrent neural network (RNN) variants Simple RNN, LSTM, and GRU on the task of dysphonia classification using the Uncommon Voice dataset. The models were trained and evaluated using standard metrics including accuracy, precision, recall, and $F_1$ score. Table 2 summarizes the performance metrics including accuracy, F1 score, precision, and recall for each RNN variant evaluated in our study.

| 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 |

Table 2. Results

The Simple RNN variant achieved an accuracy of 0.77, with precision, recall, and F1 score values of 0.5949, 0.7713, and 0.6717, respectively. Despite its simpler architecture, the Simple RNN demonstrated moderate performance in classifying dysphonia from voice recordings. This is due to the lower complexity of the RNN architecture with fewer parameters and simpler calculations. Although easier to implement and faster to train, Simple RNNs suffer from the vanishing gradient problem, limiting their ability to 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. Precision value was 0.9408 and recall was 0.9390 , indicating that the LSTM correctly identified dysphonia instances with high precision and sensitivity. The $F_1$ 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 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. Recall was 0.9391, showing its ability to capture a high proportion of positive instances in the dataset. The F1 score of 0.9372 underscores the GRU model's effectiveness in achieving both high precision and recall, providing a comprehensive evaluation of its performance. The GRU's efficiency in training and its ability to handle sequential data contributed to its competitive performance alongside the LSTM. These results highlight the comparative strengths of LSTM and GRU models over the Simple RNN in dysphonia classification using voice recordings. The higher accuracy and robustness of LSTM and GRU variants underscore their suitability for applications requiring precise classification of voice disorders. The LSTM's capability to learn and retain long-term dependencies in sequential data, along with the GRU's efficiency in training and handling sequential data, contribute to their superior performance in this task. Future research could explore further enhancements in model architectures or feature engineering to improve classification accuracy and generalization across different datasets and conditions.

## 5. Conclusion and Future Work

In this study, we evaluated three recurrent neural network (RNN) variants Simple RNN, LSTM, and GRU on dysphonia classification using the Uncommon Voice dataset. Our findings demonstrate that LSTM and GRU models outperform the Simple RNN in terms of accuracy, precision, recall, and F1 score. Specifically, LSTM achieved the highest performance with an accuracy of 0.93, followed closely by GRU with an accuracy of 0.94. These results underscore the effectiveness of LSTM and GRU architectures in capturing temporal dependencies with in voice recordings, essential for accurate dysphonia classification. The study also highlights the importance of model selection and architecture in achieving robust performance in voice disorder identification tasks. Future studies could delve deeper into the activation sequences of GRU variants, exploring specific configurations such as Candidate + Reset + Update + Forget + Activation + Output and Candidate + Update + Forget + Activation + Reset + Output + Activation to gain insights into how different gate sequences impact model performance and learning dynamics. Further comparative analyses among various GRU variants, including those with modified gate structures or activation functions, can expand our understanding of their capabilities and limitations in dysphonia classification tasks. This exploration could lead to the development of optimized GRU architectures tailored specifically for voice disorder identification.

## 6. Acknowledgment

This work was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme under Grant FRGS/1/2020/ICT06/UMS/02/1.

## References

[1] Moore, M., Papreja, P., Saxon, M., Berisha, V., & Panchanathan, (2020, October). *UncommonVoice: A crowdsourced dataset of dysphonic speech*. In *Interspeech* (pp. 2532–2536).

[2] Syed, S. A., Rashid, M., Hussain, S.,  Zahid, H. (2021). Comparative analysis of CNN and RNN for voice pathology detection. *BioMed Research International*, 2021, 6635964.

[3] 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.

[4] Graves, A. (2014). *Long short-term memory networks with residual connections for speech recognition*.

[5] Sztahó, D., Gábor, K., Gábriel, T. (2021). Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*.

[6] Merzougui, N., Korba, M. C. A., Amara, F. (2024). Diagnosing spasmodic dysphonia with the power of AI. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*.

[7] da Silva Moura, R., Maia, J. M., Dajer, M. E. (2024). Detection and classification of categories of dysphonia using convolutional neural network. In *IFMBE Proceedings* (pp. 599–610). Cham: Springer Nature Switzerland.

[8] Sindhu, I., Sainin, M. S. (2024). Automatic speech and voice disorder detection using deep learning – A systematic literature review. *IEEE Access*.

[9] Schmidhuber, J. (2000). Neural networks for compressing sequences and applications in text and speech

recognition. In *Field Guide to Neurocomputing* (pp. 917–921). Amsterdam, Netherlands: Elsevier.

[10] Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., Shih, M.-H. (2022). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. *Healthcare (Basel)*, *10*(10), 1956.

[11] 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.

[12] Shewalkar, A. (2018). *Comparison of RNN, LSTM and GRU on speech recognition data*.

[13] Khandelwal, S., Lecouteux, B., Besacier, L. (2016). *Comparing GRU and LSTM for automatic speech recognition* (Master's thesis, LIG).

[14] Hadjaidji, E., Korba, M. C. A., Khelil, K. (2021). Spasmodic dysphonia detection using machine learning classifiers. In: *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*.

[15] Rivera, M. A. (2023). Automatic identification of dysphonias using machine learning algorithms. *Applied Computer Science*, *19*, 14–25.