Fusion-Net: A Modality-Specific Weighted Ensemble of GRU Models
For Voice-Based COVID-19 Classification

David Jeffrey Rajat R Prabhu Steve Davis Abdul Hadi
National Institute of Technology, Calicut

Abstract

Recurrent Neural Networks (RNNs) have been instrumental in advancing sequential data analysis, especially in voice-based prediction tasks, due to their ability to model temporal relationships between the current input x_t and the previous hidden state h_{t-1}. However, vanilla RNNs often struggle to retain long-term dependencies because of vanishing and exploding gradients. To address this, architectures such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) introduce gating mechanisms that effectively regulate information flow and maintain context over extended sequences. In this work, we develop custom sequential classifiers based on vanilla RNNs, LSTMs, and GRUs to classify COVID-19 patients using voice data. Building on these models, we propose a new architecture called Fusion-Net, which employs a weighted ensemble approach to combine the strengths of individual classifiers and produce a more reliable prediction. The performance of all models is compared using Accuracy, Precision, Recall, and F1-Score. Experimental results show that Fusion-Net consistently achieves superior results over the baseline models, demonstrating its effectiveness and robustness for voice-based COVID-19 detection.

1 Introduction

The 21st century often referrred to as the 'Digital age' is filled with a variety of innovations and inventions which help shape the world we live in. The 'Digital age' however , like any other age had to go through one more than one kind of major turbulence-Covid19 served as a paradigm of major destruction for a large portion of the world. To contain the spread of Covid19, Testing and quarantining of the affected patients, however extensive, was done with stringent protocols to maximize safety. The standard RT-PCR test for COVID-19 was the go-to testing protocol which is accurate but expensive, slow and requires lab infrastructure.Human voice — particularly respiratory sounds such as cough, breathing, and speech — carries rich information about vocal-tract and respiratory changes induced by respiratory infections.The main idea of this project is to leverage the benefits of human voice recordings to predict the status of the patients using multiple machine learning models and optimization techniques such as the fusion net method. This approach is much cheaper and requires a basic audio file which is easily stored and accessed, for example, in a smart device such as a mobile phone. We use the coswara dataset which consists of 4 sub voice datasets each contributing to different aspects of the sound such as cough, enunciation of vowels etc.We use GRU-Gated Recurrent Unit structure for each dataset as it has proven to be the most efficient in terms of model accuracy.Each GRU model produces a single real value output indicating probability. These set of probabilities are further fed into a Fusion-Net which consists of a single layer neural network with two neurons whose outputs are used to classify whether a patient is COVID positive or not.

2 Methodology

2.1 Data thresholding

Each audio file in the dataset was trimmed using a 20 dB threshold to remove leading and trailing silence. This ensures that only the meaningful acoustic regions of the signal are preserved, improving both computational efficiency and model focus during training.

2.2 Data Augmentation

Data augmentation was done to reduce overfitting and help with class imbalance. It was applied to approximately 30% of the training samples. Two augmentation techniques employed were additive gaussian noise and time-stretching. Low-amplitude white gaussian noise was added to simulate real-world recording conditions and background disturbances. Time stretching involves changing the temporal dimension of the signal i.e. the speed of the audio slightly without altering its pitch. Small random stretch factors were applied to the signals to simulate natural variations in speaking rate. These augmentations ensure that the model remains invariant to minor acoustic and temporal perturbations in the input.

2.3 Feature Extraction

To extract meaningful representations from the audio signals, each sample was converted into a Mel spectrogram,

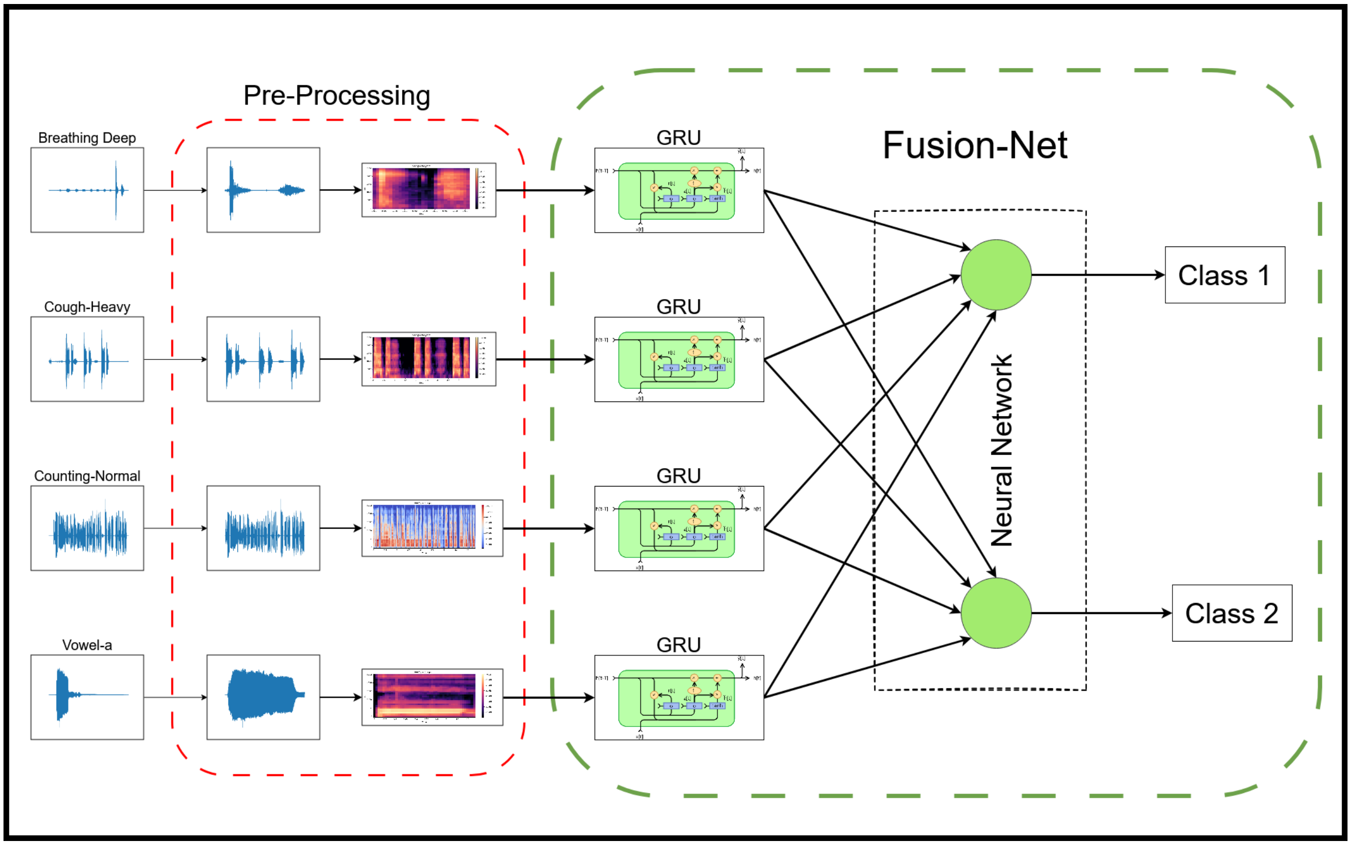

Figure 1: Architecture of the proposed Fusion-Net model combining multiple GRU networks and a final dense classifier.

which captures the time–frequency characteristics of the signal. The Mel scale was specifically used because it reflects human auditory perception, providing finer resolution at lower frequencies and compressing higher frequencies, making the representation more perceptually relevant and robust to variations across speakers and recording conditions. Additionally, the logarithm of the Mel spectrogram amplitudes was computed to approximate perceived loudness and stabilize the dynamic range. To capture temporal dynamics, the first-order (Δ) and second-order (Δ^2) derivatives of the log-Mel spectrogram were calculated, representing the rate of change and acceleration of spectral features over time. These derivatives, combined with the static log-Mel spectrogram, were used as enriched features for training the models.

2.4 Weighted ensemble

Four distinct Gated Recurrent Unit (GRU) models were trained independently, each on a different type of voice data: breathing-deep, cough-heavy, counting-normal, and vowel-a. Each model produced probability scores corresponding to its respective input type. These probabilistic outputs were then concatenated and fed into a single-layer feedforward neural network, which acts as a weighted ensemble, learning to combine and aggregate the predictions from the individual GRU models to produce the final classification result.

3 Experimental setting

3.1 Dataset

The dataset used for training was obtained from the official Coswara Project GitHub repository, developed by the Indian Institute of Science (IISc), Bangalore. It contains audio recordings of individuals performing various respiratory and speech-related tasks such as deep breathing, coughing, and vowel pronunciation. Each recording is accompanied by metadata including the participant's age, sex, and health condition (i.e., COVID-positive, negative, or not identified). The original dataset is organized into multiple folders corresponding to different dates of data collection. Each date folder contains subfolders named after participant IDs, and each participant folder includes several audio files (.wav) corresponding to different sound types (e.g., breathing-deep, breathing-shallow, cough-heavy, counting-normal, vowel-a, vowel-e). To prepare the data for model training, the dataset was restructured by sound type, creating top-level folders for each type of recording. Within each sound-type folder, recordings were further grouped based on their COVID-

19 status into three categories: Positive, Negative, and Uncertain. For improving model reliability, data points labeled as Uncertain were excluded from training. Finally, four sound categories were selected for experimentation and model development: breathing-deep, cough-heavy, counting-normal, vowel-a. This restructuring enabled targeted model training on specific sound types.

3.2 Model

In this study, we employ Gated Recurrent Units (GRUs) to train on each of the four sub-categories present in the dataset. Each GRU-based network consists of two GRU layers with an input dimension of 90 and a hidden dimension of 128. A dropout rate of 40% is applied to prevent overfitting. Owing to the relatively small dataset, a bidirectional GRU architecture was not adopted, as it may increase model complexity and lead to overfitting. The recurrent layers are followed by a fully connected hidden layer comprising 64 neurons, and finally, an output layer with two neurons for binary classification. The Rectified Linear Unit (ReLU) activation function is employed for non-linearity.

During training, the model is optimized using the Adam optimizer with a learning rate (α) of $3 * 10^{-}4$ and a weight decay of $1 * 10^{-}5$.The loss function used is CrossEntropyLoss. To prevent overfitting and enhance generalization, a learning rate scheduler—ReduceLROnPlateau—is utilized with a patience of five epochs. Each GRU model is trained for 20 epochs under these conditions. Similar architectural and training configurations are used for the LSTM and vanilla RNN models for a fair comparison.

For the proposed Fusion-Net, the architecture consists of a single output layer with two neurons. The input to this layer comprises the concatenated probability score vectors obtained from the individually trained GRU models corresponding to each voice sub-category. The Fusion-Net is trained using the CrossEntropyLoss function and the Adam optimizer with a learning rate of $1 * 10^{-}3$ for 10 epochs

4 Evaluation and Comparison

The performance of all four models—RNN, LSTM, GRU, and the proposed Fusion-Net—was assessed using the standard evaluation metrics of Accuracy, Precision, Recall, and F1-Score. These metrics were computed from the confusion matrix, which gives a detailed picture of how well each model distinguishes between COVID-positive and COVID-negative samples. To make the comparison clearer, the confusion matrices were visualized as heatmaps, allowing easy identification of classification trends and misclassifications across models. The quantitative results for each classifier are summarized in Table [1]. Among all models, the proposed Fusion-Net achieved the highest accuracy and F1-score, showing a strong balance between sensitivity and specificity. The GRU-based classifiers also produced solid results, outperforming both the vanilla RNN and LSTM models. This demonstrates the advantage of using gated architectures when dealing with sequential voice data. From the classification reports, it is evident that Fusion-Net attains higher recall compared to the other models, indicating better reliability in identifying COVID-positive cases. At the same time, it maintains strong precision, which helps reduce false positives. Overall, these findings suggest that the ensemble-based design of Fusion-Net effectively combines the strengths of individual GRU classifiers and offers improved generalization on unseen voice samples.

Table 1: Comparison of classifier evaluation metrics — Accuracy, Precision, Recall, and F1-Score — for vanilla RNN, LSTM, GRU, and the proposed Fusion-Net models on the voice-based COVID-19 dataset.

2*Model	**Precision**		**Recall**		**F1-Score**		**Acc.**
	Pos	Neg	Pos	Neg	Pos	Neg	
RNN	0.525	0.79	0.43	0.85	0.465	0.815	0.73
LSTM	0.5875	0.8125	0.515	0.845	0.5375	0.8275	0.7525
GRU	0.6225	0.82	0.515	0.8775	0.5575	0.8475	0.77
Fusion-Net	**0.87**	**0.86**	**0.62**	**0.96**	**0.73**	**0.91**	**0.87**

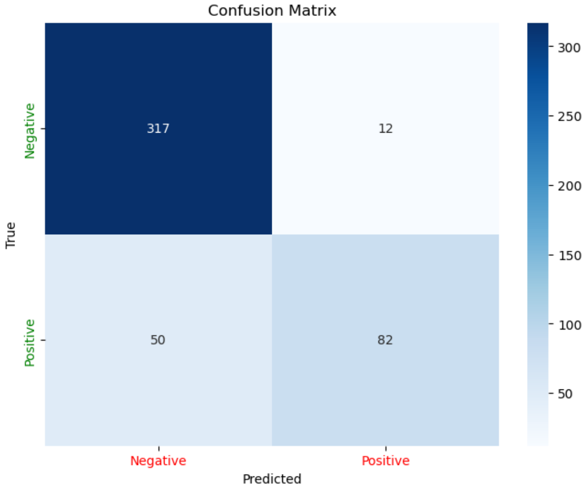

Figure 2: Confusion matrix of the proposed Fusion-Net model showing classification performance on the voice-based COVID-19 dataset. The matrix illustrates the distribution of correctly and incorrectly classified COVID-positive and COVID-negative samples.

5 Complexity Analysis

Computational and memory complexity are critical factors in sequential models, particularly when considering power consumption and deployment efficiency. In recurrent networks, the most computationally expensive operations are the **multiply–accumulate (MAC) operations**, which dominate both time and resource requirements.

We present a detailed analysis for Vanilla RNNs, LSTMs, GRUs, and the proposed Fusion-Net.

Symbol	Description	Dimension
W_{xh}	Input Weight Matrix	$\mathbb{R}^{H\times D}$
W_{hh}	Hidden Weight Matrix	$\mathbb{R}^{H\times H}$
W_{hy}	Output Weight Matrix	$\mathbb{R}^{O\times H}$
x_t	Input	$\mathbb{R}^{D}$
h_t	Hidden State	$\mathbb{R}^{H}$
y_t	Output	$\mathbb{R}^{O}$

Table 2: Summary of notations and corresponding dimensional representations for parameters and variables in RNN-based architectures.

Symbol	Description	Value
D	Input Feature Dimension	90
H	Hidden State (Layer) Dimension	128
O	Output Dimension	128
T	Number of Time Steps	T
L	Number of RNN Layers	2
F	Fully Connected Layer Neurons	64
C	Number of Classes	2
G	Number of GRUs used	4

Table 3: Model parameters and their symbolic representations used for complexity computation.

5.1 Vanilla RNN

The update equations for a vanilla RNN are:

$$h_t = g(W_{xh}X_t + W_{hh}h_{t-1} + b_x), \quad y_t = W_{hy}h_t + b_y$$

where $g(\cdot)$ denotes the activation function (commonly tanh or ReLU).

Time Complexity:

- Input layer MACs: $D \times H$
- Hidden layer MACs (feedback): $H \times H$
- For T timesteps and L layers, total MACs per sample:

$$\text{MACs} = LTH(D + H) + OH$$

- **Time complexity:** $\mathcal{O}(LTH(D + H))$

Space Complexity:

- Total parameters (including biases):

$$L(DH + H^2 + H) + HO + O$$

- Runtime memory to store hidden states: $\mathcal{O}(THL)$
- **Space complexity:** $\mathcal{O}(L(DH+H^2+H))+\mathcal{O}(THL)$

5.2 LSTM

The LSTM cell uses the following equations:

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)\\
\tilde{c}_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t\\
h_t &= o_t \odot \tanh(c_t)\\
y_t &= W_{hy}h_t + b_y
\end{aligned}$$

Time Complexity:

- MAC operations:

$$\text{MACs} = 4LTH(D + H) + OH$$

- **Time complexity:** $\mathcal{O}(4LTH(D + H))$

Space Complexity:

- Total parameters (including biases):

$$4L(DH + H^2 + H) + HO + O$$

- Runtime memory: $\mathcal{O}(4THL)$
- **Space complexity:** $\mathcal{O}(4L(DH + H^2 + H)) + \mathcal{O}(4THL)$

5.3 GRU

The GRU cell is defined by:

$$\begin{aligned}
z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z)\\
r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r)\\
\hat{h}_t &= \tanh(W_{xh}x_t + W_{hh}(r_t \odot h_{t-1}) + b_h)\\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t\\
y_t &= W_{hy}h_t + b_y
\end{aligned}$$

Time Complexity:

- MAC operations:

$$\text{MACs} = 3LTH(D + H) + OH$$

- **Time complexity:** $\mathcal{O}(3LTH(D + H))$

Space Complexity:

- Total parameters (including biases):

$$3L(DH + H^2 + H) + HO + O$$

- Runtime memory: $\mathcal{O}(3THL)$
- **Space complexity:** $\mathcal{O}(3L(DH + H^2 + H)) + \mathcal{O}(3THL)$

5.4 Fusion-Net

Fusion-Net combines G GRU networks with a single-layer neural network for final classification.

Time Complexity:

- MAC operations:

$$\text{MACs} = G\big(3LTH(D+H)+OH+OF+FC\big)+GC^2$$

Space Complexity:

- GRU parameters

$$G\times(3L(DH+H^2+H)+HO+O+OF+F+FC+C)$$

- Neural network parameters

$$G \times C^2 + C$$

- Total parameters

$$G\times(3L(DH+H^2+H)+HO+O+OF+F+FC+C) + (G \times C^2 + C) \tag{1}$$

- Space Complexity

$$\text{Space complexity} \approx \boldsymbol{\mathcal{O}(G \times 3L(DH + H^2 + H))}$$

Model	**MAC Operations**	**Parameters**
Vanilla RNN	80,512	80,962
LSTM	2,47,936	2,49,154
GRU	1,92,128	1,93,090
Fusion-Net	7,68,528	7,72,378

Table 4: Comparison of Computational Complexity and Model Parameters Across Architectures

6 Limitations and Future Work

While the proposed Fusion-Net achieves notable improvements in predictive performance, these gains are accompanied by increased computational overhead. The model exhibits substantially higher multiply–accumulate (MAC) operations and parameter counts compared to conventional GRU and LSTM architectures, leading to longer training times, greater memory requirements, and higher inference latency. Such complexity may restrict its applicability in resource-constrained or real-time environments.

Future work will focus on enhancing the computational efficiency of Fusion-Net through techniques such as model pruning, quantization, and knowledge distillation, aiming to preserve its predictive accuracy while reducing computational and memory demands. Additionally, exploring lightweight architectural variants or hardware-aware optimization strategies could further improve its deployment feasibility on edge and mobile platforms. Another promising direction involves expanding the dataset with more diverse voice samples to improve model generalization across demographics and recording conditions. Furthermore, integrating advanced temporal attention or transformer-based fusion mechanisms could enhance the interpretability and adaptability of Fusion-Net in future iterations.

7 Conclusion

This study presented a comparative analysis of sequential deep learning architectures—vanilla RNN, LSTM, and GRU—alongside a custom weighted ensemble-based model, Fusion-Net, for classifying COVID-19 patients using voice data. The experimental results reaffirm the limitations of traditional RNNs in capturing long-term temporal dependencies, while gated architectures such as LSTMs and GRUs demonstrated notable improvements in both stability and performance. Building upon these insights, the proposed Fusion-Net effectively harnessed the complementary strengths of multiple GRU-based classifiers through a weighted ensemble mechanism, thereby enhancing robustness and generalization.

Among all evaluated models, Fusion-Net achieved the highest accuracy and F1-score, reflecting a strong balance between sensitivity and specificity. The model also exhibited superior recall, indicating enhanced reliability in identifying COVID-positive cases, while maintaining high precision to minimize false positives.

References

[1] C. Brown *et al.*, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," *Proc. ACM SIGKDD*, 2020.

[2] V. Despotovic, L. Ismail, and J. Cornet, "Detection of COVID-19 from voice, cough and breathing patterns: A systematic review," *J. Biomed. Inform.*, vol. 118, 2021.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, 2018.

[6] N. Sharma *et al.*, "Coswara – A database of breathing, cough, and voice sounds for COVID-19 diagnosis," *arXiv preprint arXiv:2005.10548*, 2020.