Deep recurrent neural networks with attention mechanisms for respiratory anomaly classification.

WALL, C., ZHANG, L., YU, Y. and MISTRY, K.

2021

Deep Recurrent Neural Networks with Attention Mechanisms for Respiratory Anomaly Classification

Conor Wall
Department of Computer and Information Sciences
Faculty of Engineering and Environment, Northumbria University
Newcastle, UK, NE1 8ST
cwall1996@hotmail.com

Li Zhang
National Subsea Centre
School of Computing
Robert Gordon University
Aberdeen, UK, AB10 7AQ
l.zhang7@rgu.ac.uk

Yonghong Yu
College of Tongda
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China
yuyh@njupt.edu.cn

Kamlesh Mistry
Department of Computer and Information Sciences
Faculty of Engineering and Environment, Northumbria University
Newcastle, UK, NE1 8ST
k.mistry@northumbria.ac.uk

***Abstract*—In recent years, a variety of deep learning techniques and methods have been adopted to provide AI solutions to issues within the medical field, with one specific area being audio-based classification of medical datasets. This research aims to create a novel deep learning architecture for this purpose, with a variety of different layer structures implemented for undertaking audio classification. Specifically, bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Units (GRU) networks in conjunction with an attention mechanism, are implemented in this research for chronic and non-chronic lung disease and COVID-19 diagnosis. We employ two audio datasets, i.e. the Respiratory Sound and the Coswara datasets, to evaluate the proposed model architectures pertaining to lung disease classification. The Respiratory Sound Database contains audio data with respect to lung conditions such as Chronic Obstructive Pulmonary Disease (COPD) and asthma, while the Coswara dataset contains coughing audio samples associated with COVID-19. After a comprehensive evaluation and experimentation process, as the most performant architecture, the proposed attention BiLSTM network (A-BiLSTM) achieves accuracy rates of 96.2% and 96.8% for the Respiratory Sound and the Coswara datasets, respectively. Our research indicates that the implementation of the BiLSTM and attention mechanism was effective in improving performance for undertaking audio classification with respect to various lung condition diagnoses.**

Keywords—deep learning, Long Short-Term Memory, audio classification, lung disease, COVID, bidirectional Recurrent Neural Network, attention mechanism

I. INTRODUCTION

Deep learning methods are now one of the most prominent methods in computing today with respect to tasks such as audio, video, and image classification [1]. In the field of medical imaging, a large number of deep learning methods have been demonstrated in a variety of studies. Some example deployments pertaining to medical image classification include melanoma identification, diabetic retinopathy screening, and blood cancer detection. Such existing deep learning research has resulted in significant performance enhancement with respect to medical diagnosis [2]. Besides the above, there have also been studies with respect to the audio classification of medical datasets. As an example, Convolutional Neural Networks (CNNs) and traditional machine learning methods, such as Support Vector Machine (SVM), have been adopted in [3] for lung condition diagnosis using audio datasets. As one specific type of Recurrent Neural Network (RNN), the Long Short-Term Memory (LSTM) network is widely adopted for time series forecasting [4]. In a recent research study, Kumar et al. [5] employed an LSTM model for heartbeat audio classification, which yielded an accuracy rate of 80%, outperforming all other machine learning methods utilised in their experiments [5]. There are also a variety of other existing studies that indicated the efficiency of deep learning methods pertaining to audio classification tasks [6, 7, 8, 9].

Motivated by aforementioned existing studies, in this research, we explore the use of deep learning models, specifically RNN architectures with attention mechanisms for the classification of medical audio datasets for chronic and non-chronic lung diseases, as well as COVID-19 diagnosis. We evaluate the proposed attention RNN models using both the Respiratory Sound [6] and the Coswara datasets [7]. Specifically, the Respiratory Sound Database contains audio data with respect to a total of six lung conditions such as Chronic Obstructive Pulmonary Disease (COPD) and asthma, while the Coswara dataset contains coughing audio samples associated with COVID-19. The empirical results indicate that the proposed models with attention mechanisms outperform other deep learning architectures for diverse lung condition classification.

II. RELATED WORK

A. Related Studies for Audio Classification

There have been several existing studies that explored the combination of CNN and LSTM for audio classification. In particular, the investigation of classifying music genres has been intensively studied with impressive results using deep learning methods. As an example, Choi et al. [8] employed a

Convolutional Recurrent Neural Network (CRNN), which is described as a CNN model with the last layers replaced with an RNN network. The Million Song Dataset, consisting of numerous song clips, has been employed for model evaluation for the classification of categories such as genre, mood, era, and instrument [9]. Their CRNN model was composed of six layers which included four conv2d layers and two RNN layers. To utilise the dataset for model evaluation, features were extracted from the audio files using the python package Librosa. The features extracted from the audio files are known as Mel-Frequency Cepstral Coefficients (MFCC), which are the logarithmic measure of the Mel magnitude spectrum and contain sufficient discriminating properties. This in particular makes them efficient assets for classifying audio datasets [10]. The empirical results of their studies indicated that their proposed CRNN model outperformed all three other baseline CNN models consistently for audio classification. In particular, the CRNN model outperformed the CNN model with 5 convolutional layers and 2 fully-connected layers on 44 tags out of the 50 tags in the dataset, based on the AUC-ROC (Area Under Curve-Receiver Operating Characteristic Curve) metric, pertaining to music tagging. However on the other hand, their proposed CRNN model had the highest number of model parameters and was computationally costly.

Zheng et al [11] demonstrated another CRNN model for Gastrointestinal (GI) sound event detection. Their work employed a gastrointestinal sound dataset that includes 6 different types of body sounds, i.e. bowel sound, speech, snore, cough, groan, and rub. As in the existing studies, to utilise the audio files, MFCC features were extracted using the python package Librosa. Their proposed CRNN model was made up of a 5-layer CNN network, followed by a bidirectional Gated Recurrent Unit (BiGRU) layer and the fully connected layers. Their work achieved promising results with the mean F1 score of 81.06% for the detection of the aforementioned 6 categories of sounds, with two of the classes, i.e. speech and snore, yielding F1 scores over 90%.

B. Bidirectional RNN Architectures

In the above-mentioned existing studies, the concept of a BiGRU was utilised as a part of the implemented model. With the goal of this research to implement an RNN model in conjunction with attention mechanisms, the concept of using a bidirectional RNN is worth exploring further. While a GRU is not an LSTM, it is very similar in functionality and design, with the main difference being that the GRU combines the "forget" and "input" gates into an "update" gate, as well as adding a "reset" gate. This results in the GRU model having fewer parameters and generally a simpler architecture than that of an LSTM network [12].

However, one key distinction of our research is the use of bidirectional design for the LSTM and GRU models. A bidirectional RNN model has two RNN layers of the same type, for example, having two LSTM layers. These two layers ensure that the input features can be processed in both forward and backward directions. This enables the model to better obtain the relations among elements in the input sequence by using the information in both forward and backward directions [13].

In addition, Chen and Li [14] demonstrated a CNN-BiLSTM model for the classification of emotions embedded in music. The dataset adopted in their work consisted of 2000 audio song samples from the Last.fm tag subset of the Million Song Dataset [9]. The dataset consisted of 500 song samples for each of the following emotion classes, i.e. anger, sadness, relaxation, and sadness. Like aforementioned studies, MFCC features were extracted from the audio files and adopted for model training. Their studies indicated that their proposed CNN-BiLSTM method achieved an average accuracy rate of 68% across the four classes, while the other baseline models, i.e. CNN-LSTM, CNN, and LSTM, achieved 63%, 59%, and 50% respectively for music emotion classification. This demonstrates again that the use of BiLSTM layer architectures can potentially increase classification performance. As the use of bidirectional RNN methods was shown to be advantageous in [11] and [14], the concept has been further explored in our research.

C. Attention Mechanisms

Another concept that has been explored intensively in various studies as part of the RNN architectures is the attention mechanism, which has shown very positive results in areas such as speech recognition and natural language processing (NLP). The attention mechanism provides an adaptive ability to learn the relationship of each of the input features at several time steps to predict the current time step [15].

In the work of Zhang et al. [16], a convolutional RNN architecture with an attention mechanism, namely ACRNN, was proposed. Attention for both CNN and RNN layers was investigated. In particular, their work focused on being both determining the effectiveness of the attention mechanism as well as the position that the attention mechanism should reside in within the model. Their work was evaluated using the environmental audio datasets, i.e. ESC-50 and ESC-10, which consist of 50 and 10 classes respectively [17]. The empirical results indicated that the attention mechanism provided a significant increase in accuracy, with an over 2% increase for both datasets. It was also found that the attention mechanism was best suited for increasing classification accuracy in layers 2 and 10 in their CRNN network. These discoveries indicated that an attention mechanism implemented within a deep CRNN model would be beneficial for undertaking audio classification and is worth further investigation.

III. The Proposed Deep Networks with Attention Mechanisms

In this research, we propose two bidirectional LSTM and GRU networks incorporated with attention mechanisms, namely A-BiLSTM and A-BiGRU, for chronic and non-chronic lung conditions and COVID-19 diagnosis. We introduce the key procedures such as the feature extraction from audio inputs and the proposed deep learning models in detail below.

A. Feature Extraction

As indicated in the aforementioned existing studies, the first step is to extract the features from audio inputs. The method of choice includes the extraction of the MFCC features, which again can be achieved through the use of the python package Librosa [18].

There are several aspects that need to be taken into account in the pre-processing stage. The first of which is to decide how to ensure that the model has enough input features, which can be achieved by splitting the audio files into segments. Depending on the sample rate and the length of each of the audio clips, which need to be determined, the audio input can then be split into segments.

Following the splitting stage, all of the MFCC features are extracted from each of the segments and then appended to a dictionary with its class label. To achieve this, certain variables need to be decided upon such as the value for the Fast Fourier Transform (FFT) algorithm and the hop length. The FFT algorithm is typically used to convert a signal from its original domain, which in this case is time, to a representation in the frequency domain. In the context of MFCC, FFT is applied to every frame to calculate the frequency spectrum. This is conducted through the process called the Short-Time Fourier-Transform (STFT), from which the power spectrum is then calculated. Once the power spectrum is calculated, then triangular filters applied on the Mel-scale are applied to the power spectrum to extract frequency bands. Using these frequency bands, the Mel frequency is computed through the use of the formula below [20].

$$Mel(f) = 1127 \text{ x } \ln\left(1 + \frac{f}{700}\right) \qquad (1)$$

The formula in Equation (1) in particular converts the audio input to the Mel frequency in hertz i.e $Mel(f)$. First, a setting of 1127 is calculated by taking the natural logarithm (ln) and the corner frequency of 700 hertz, which is typically between 600 and 1000 hertz for this type of formula. This is then multiplied by the natural logarithm (ln), where a constant value of 1 pluses the frequency in hertz (*f*) being divided by the corner frequency of 700.

The values of the hop length combined with the FFT algorithm determine how many frames are taken from each segment. The default numbers for both of these variables are 2048 for FFT and 512 for hop length, which for simplicity, will be used for this research. Following the extraction of the MFCC feature from the audio files, they are each appended onto a JSON file which is used as the input file for training the model architectures.

B. The Proposed Model Architectures

In this research, we propose bidirectional RNN models with an attention mechanism for audio classification, i.e., chronic and non-chronic lung conditions and COVID-19 identification via breathing, coughing, and voice recordings. The two specific types of RNN networks chosen for model construction are LSTM and GRU. Determining the specific structure of the networks involves rigorous testing of the settings of each of the layers' parameters, as well as the hyperparameters during the training process. Table 1 below describes the architectures of the proposed BiLSTM and BiGRU models with the attention mechanisms.

TABLE I. A-BILSTM AND A-BIGRU MODEL ARCHITECTURES

Layer	Layer Description	Unit Setting
1	BiLSTM or BiGRU	512
2	Attention Mechanism	N/A
3	LSTM or GRU	512
4	Dense (Relu)	256
5	Dropout	0.5
6	Dense	128
Fully Connected	Dense (Softmax)	6/2

As shown in Tables I, the two model architectures of A-BiLSTM and A-BiGRU have the same structures, with the only differing aspect being the type of the RNN network, i.e., LSTM or GRU, implemented in layers 1 and 3. The choice of implementing the attention mechanism in the second layer of each model was influenced by the suggestion in [16], which demonstrated that the attention mechanism was best suited to increase the model accuracy by being on layer 2 or 10 of the network. With the choice of a dense layer being the final connected layer of the two architectures, therefore the only remaining option was to implement the mechanism on layer 2.

In addition, the first layer of the model was decided to be the aforementioned BiLSTM or BiGRU layer, with the hidden neuron units set as 512. Owing to the layer being bidirectional, the number of hidden units is then doubled. The number of hidden units was determined based on rigorous testing which involved experimenting with different numbers of neurons.

Following the attention layer, an unidirectional LSTM or GRU layer is then implemented, with the choice of units being set as 512 as it is half the value of the first layer. Following the third layer, a regular dense layer with the activation function 'relu' is found to be the most effective. The next layer is a dropout layer, which is to reduce the amount of overfitting that may occur during the training of neural networks [21].

The two final layers implemented are the two dense layers, the first being 128 units, and the final being the fully connected layer consisting of 6 or 2 units (i.e., the number of classes), depending on the expected number of classes being outputted in the employed test dataset. In addition, the specific classes for both datasets are explained in the next section.

TABLE II. BILSTM AND BIGRU MODEL ARCHITECTURES

Layer	Layer Description	Unit Setting
1	BiLSTM or BiGRU	512
2	LSTM or GRU	512
3	Dense (Relu)	256
4	Dropout	0.5
5	Dense	128
Fully Connected	Dense (Softmax)	6/2

The fully connected layer also has the activation function 'softmax'. To determine the effectiveness of the attention layer,

testing also involved training both model architectures without the attention mechanism implemented, as shown in Table II.

C. Model Training

As mentioned previously, the training process was rigorous to optimise the layers of the models and the various training hyperparameters.

The first choice to be made pertaining to the training and test processes is determining the train, validation, and test split of the dataset. Several divisions were tested, but what was conclusively chosen was the splits shown in Table III below.

TABLE III. TRAINING, VALIDATION, TEST SPLITS FOR EACH TEST DATASET

Datasets			Train (%)	Validation (%)	Test (%)
Respiratory Sound Database	75	12.5	12.5		
The Coswara Dataset	80	10	10		

These splits were found to be the most effective for optimising the model performance and accuracy for the respective datasets.

Due to the nature of the two networks, i.e. A-BiLSTM and A-BiGRU, being fundamentally different, with the GRU network being a simpler variation of the LSTM, the hyperparameters identified to achieve optimal performance between the two networks differed a fair amount. Tables IV-V below demonstrate the identified optimal model settings.

TABLE IV. A-BILSTM AND BILSTM MODEL HYPERPARAMETERS

Hyperparameters	Setting
Epochs	150
Learning Rate	0.00001
Batch Size	64

TABLE V. A-BIGRU AND BIGRU MODEL HYPERPARAMETERS

Hyperparameters	Setting
Epochs	20
Learning Rate	0.01
Batch Size	64

As shown in Tables IV-V, the training process for the GRU models required far fewer epochs, as well as a comparatively larger learning rate, to achieve optimal performance, which reflects the comparatively simple nature of the networks themselves.

Moreover, Figs. 1-2 are examples of the training and validation losses for the Coswara dataset, with respect to A-BiLSTM and A-BiGRU, respectively. As demonstrated in Figs. 1-2, no overfitting occurred during the training process which ensures that all models are performing to the best of their capability. We discuss the evaluation details in the following section.

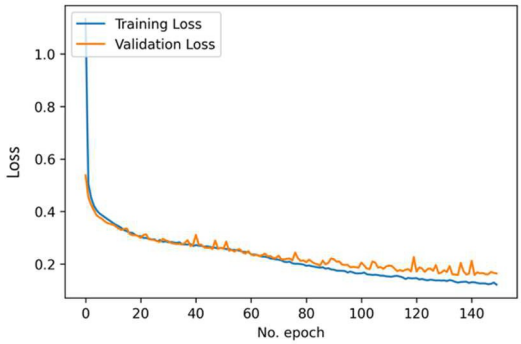

Fig.1. The training and validation losses for the A-BiLSTM model with respect to the Respiratory Sound Database

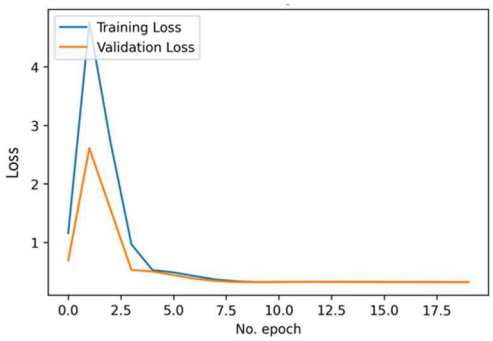

Fig. 2. The training and validation losses for the A-BiGRU model with respect to the Coswara dataset

IV. EVALUATION

To ensure a comprehensive model evaluation, both the Respiratory Sound Database and the Coswara dataset are used in our experiments to give a good indication of the proficiency and effectiveness of the proposed models.

The Respiratory Sound Database was chosen for investigation in this research. This is mainly owing to the fact that three out of the top ten leading causes of death globally are respiratory diseases [21]. With early detection being crucial for preventing deaths of such diseases, providing a more convenient and accurate method of diagnosis could offer a vital tool to medical professionals working in the respiratory field [22]. Specifically, the dataset contains 920 recordings from 126 subjects whose conditions include Healthy, Upper Respiratory Tract Infections (URTI), Chronic Obstructive Pulmonary Disease (COPD), Bronchiolitis, Pneumonia and Bronchiectasis. In other words, a total of six lung conditions are taken into account for model evaluation. All the 920 audio recordings have been employed in our experiments. With there being numerous classes to classify from and the sounds being recorded from a range of stethoscopes, to achieve high

accuracy would require an extremely proficient model, which constitutes a challenging scenario for audio classification.

We also employ the Coswara dataset to test model efficiency. Owing to the contents of the dataset being strictly COVID-19 related, the obvious primary reason of selecting the dataset would be to research and provide a possible solution that could help alleviate the current worldwide pandemic.

The current most widely used method of testing for COVID-19 is the PT-PCR test, which while it is the most effective method of testing at this moment in time, it also has several issues such as cost, scalability, and the nature of the test violating social distancing [23]. Providing a more convenient, cost-effective, and scalable method of diagnosis would deliver a crucial service to allow more people to be tested daily and ultimately provide control over the pandemic.

The Coswara dataset itself is open access and consists of a growing number of respiratory audio recording classes that include coughing, breathing, and voice recordings. The subsection focussed on in this research is the cough class. This subsection can be classified into three categories, i.e. healthy subjects, subjects who have COVID-19, and subjects who have a respiratory disease that is not COVID-19.

The two classes that are included in this study are healthy subjects and subjects who have COVID-19, in other words, positive and negative cases for COVID-19 diagnosis. A total of 95 positive and 100 negative cases are employed for model evaluation.

Tables VI-VII below illustrate the performance of the four RNN models on both datasets. The results indicate that the best performing model is the A-BiLSTM network for undertaking audio classification for both datasets.

TABLE VI EXPERIMENTAL RESULTS FOR THE RESPIRATORY SOUND DATABASE

Models	Average Test Accuracy (%)
A-BiLSTM	96.2
BiLSTM	93.2
A-BiGRU	93
Bi-GRU	91.2

TABLE VII. EXPERIMENTAL RESULTS FOR THE COSWARA DATASET

Models	Average Test Accuracy (%)
A-BiLSTM	96.8
BiLSTM	94.6
A-BiGRU	94.2
Bi-GRU	92.8

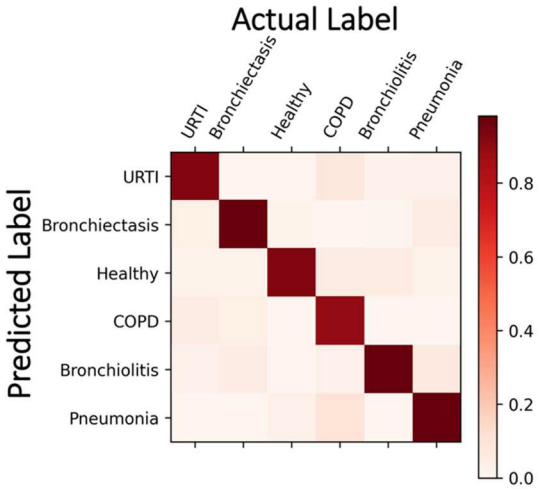

Fig.3. Confusion matrix of the proposed A-BiLSTM model for the Respiratory Sound Database.

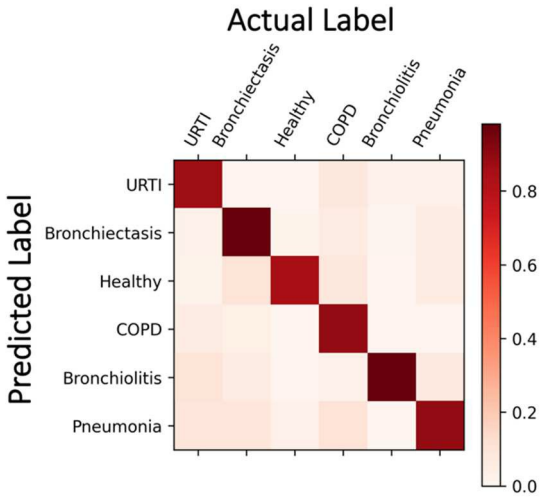

Fig.4. Confusion matrix of the proposed A-BiGRU model for the Respiratory Sound Database

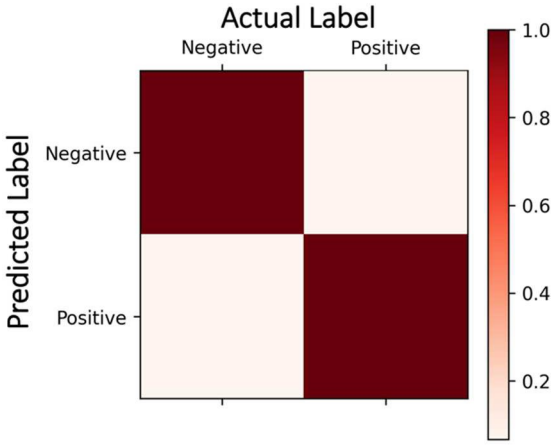

Fig.5. Confusion matrix of the proposed A-BiLSTM model for the Coswara dataset

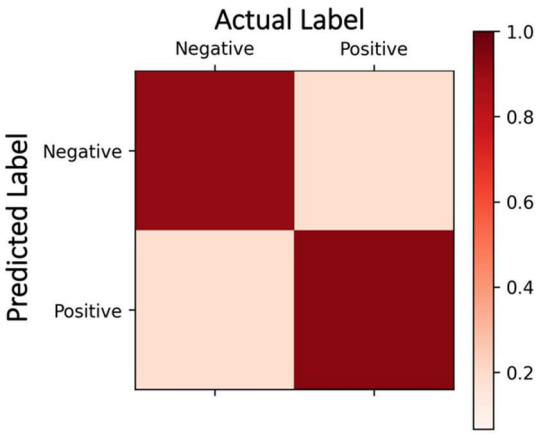

Fig.6. Confusion matric of the proposed A-BiGRU model for the Coswara dataset

Moreover, Figs. 3-6 illustrate the confusion matrices for the Respiratory Sound database and Coswara datasets for both of the both proposed models, i.e. A-BiLSTM and A-BiGRU. Tables VIII-XI below illustrate the detailed results of each class for both datasets with respect to the A-BiLSTM and A-BiGRU networks.

TABLE VIII. DETAILED RESULTS FOR A-BILSTM FOR THE RESPIRATORY SOUND DATABASE

Class	Test Accuracy (%)
URTI	94.8
Bronchiectasis	98.8
Healthy	94.3
COPD	93.3
Bronchiolitis	98.8
Pneumonia	97.2

TABLE IX. DETAILED RESULTS FOR A-BIGRU FOR THE RESPIRATORY SOUND DATABASE

Class	Test Accuracy (%)
URTI	88.8
Bronchiectasis	97.2
Healthy	87.3
COPD	93.5
Bronchiolitis	97.8
Pneumonia	93.3

TABLE X. DETAILED RESULTS FOR A-BILSTM FOR THE COSWARA DATASET

Class	Test Accuracy (%)
Positive	97.6
Negative	96

TABLE XI. DETAILED RESULTS FOR A-BIGRU FOR THE COSWARA DATASET

Class	Test Accuracy (%)
Positive	93.3
Negative	95.1

V. DISCUSSIONS

From the inspection of the empirical results, several observations can be made. The main finding is that for both datasets, as indicated in Tables VI-VII, the A-BiLSTM model was the best performing method in terms of accuracy, i.e., it obtains accuracy rates of 96.2% and 96.8% for the Respiratory Sound and the Coswara datasets, respectively. Similarly, for both datasets, the LSTM models outperform the respective GRU models for each of the experiments.

One possible explanation for both of the above observations is that because of the more complex nature of the LSTM, with it embedding more complex layer topologies, it provides the network with better capability to learn bidirectional temporal dependencies on large MFCC datasets. Moreover, as seen in Figs. 1-2, the number of epochs required to train on the Coswara dataset to achieve optimal performance was far fewer for the A-BiGRU model in comparison with those of the A-BiLSTM model.

As discussed previously, this is more likely owing to the simpler nature of the A-BiGRU model. Therefore, if the efficiency or size of the model is essential in determining the models to use, for example, in the case of using a model on a mobile device, then the GRU would be preferable.

Another key observation that can be made is with respect to the impact on performance of the implemented attention mechanism. Not only does the highest performing model for both datasets consist of the attention mechanism, but all models that have an attention mechanism implemented show higher performance than their counterparts, with the difference in test accuracy ranging from 1.8% and 3% for the Respiratory Sound Database, and 1.4% to 2.2% for the Coswara dataset.

Although the improvements are not transformative, they are notable and significant enough to be worthy of implementation, especially when considering that achieving test accuracy results of 96.8% for the Coswara dataset would not be possible without the attention mechanism being implemented. Despite that in this particular study, no models tested consist of purely unidirectional LSTM/GRU layers, we can observe the results from [4] which demonstrated that an unidirectional LSTM model tested on a heartbeat audio dataset achieved an accuracy rate of 80%.

Despite the datasets being tested not being the same as in this research, they are similar in nature as they are all audio medical datasets. Therefore, the superiority of our experimental results of around 96% on both selected datasets indicates the potential of the use of bidirectional RNN to be more suited to achieve higher performance on audio

classification, owing to the consideration of both forward and backward states simultaneously during the inference process.

Another conclusion that can be made by observing the Figs. 3-6 and Tables VIII-XI is that both the confusion matrices and the detailed experimental results illustrate how the A-BiLSTM model outperforms A-BiGRU significantly for the classification of each class with respect to both datasets.

Specifically by inspecting the results of the Respiratory Sound Database in Tables VIII-IX, the A-BiLSTM model outperforms the A-BiGRU network for the classification of nearly all the categories. In particular, the A-BiLSTM model yielded substantially better results for the categories such as Healthy (94.3%), Pneumonia (97.2%) and UTRI (94.8%), when compared with the A-BiGRU model, which yielded 87.3%, 93.3% and 88.8% for Healthy, Pneumonia and UTRI, respectively.

On the other hand, for the Coswara dataset, according to the experimental results shown in Tables X-XI, the A- BiLSTM model again outperforms the A-BiGRU model for the prediction of both positive and negative classes.

Ultimately, our conclusion is that the specific architecture implemented for the A-BiLSTM model has achieved impressive results and is certainly worth exploring and experimenting with further.

VI. Conclusion

In this research, we have implemented novel architectures of the RNN networks with attention mechanisms for the classification of the medical audio datasets, i.e., the Respiratory Sound and the Coswara datasets. The experimental results of both datasets reveal the efficiency of the proposed A-BiLSTM network among all the test methods, which consists of a BiLSTM layer, an attentional layer, an unidirectional LSTM layer, and several dense layers. The A-BiLSTM model achieves the test accuracy rates of over 96% for both datasets, which indicates that the implementation of a BiLSTM network and attention mechanism is a concept that is beneficial for improving audio classification performance.

Furthermore, the aforementioned A-BiLSTM architecture implemented in this research was shown to be highly effective, but with further experimentation with different layer and hyperparameter settings [24-32], additional improvements in performance could be made. Evolutionary algorithms [33-49] could also be exploited pertaining to the above parameter tuning as well as architecture generation processes. Moreover, it would also be beneficial to employ additional medical audio datasets to further evaluate model efficiency.

Acknowledgment

We would like to acknowledge the funding support received from the Purposeful Health Growth Accelerator project funded by Research England.

References

[1] C. Wall, F. Young, L. Zhang, E. Phillips, R. Jiang and Y. Yu, "Deep learning based melanoma diagnosis using dermoscopic images", *Developments of Artificial Intelligence Technologies in Computation and Robotics*, 2020.

[2] R. Yamashita, M. Nishio, R. Do and K. Togashi. "Convolutional neural networks: an overview and application in radiology". *Insights into Imaging*, 9(4), pp.611-629, 2018.

[3] M. Aykanat, Ö. Kılıç, B. Kurt and S. Saryal. "Classification of lung sounds using convolutional neural networks." *EURASIP Journal on Image and Video Processing*, 2017.

[4] J. Kumar, R. Goomer, and A. Singh. "Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters." *Procedia Computer Science*, 125, pp.676-682. 2018.

[5] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G. Choi and B.W. On, "Heartbeat Sound Signal Classification Using Deep Learning", *Sensors*, vol. 19, no. 21, p. 4819, 2019.

[6] B.M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques and R.P. Paiva. "A respiratory sound database for the development of automated classification". In *International Conference on Biomedical and Health Informatics* (pp. 33-37). Springer, Singapore. 2017.

[7] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S.R. Chetupalli, P.K. Ghosh, and S. Ganapathy. "Coswara-A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis", 2020. *arXiv preprint arXiv:2005.10548*.

[8] K. Choi, G. Fazekas, M. Sandler and K. Cho, "Convolutional recurrent neural networks for music classification", in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] T. Bertin-Mahieux, D.P.W. Ellis, B. Whitman and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, Florida, USA, October 24-28, 2011.

[10] S. Zahid, F. Hussain, M. Rashid, M. Yousaf and H. Habib, "Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods", *Mathematical Problems in Engineering*, vol. 2015, pp. 1-11, 2015.

[11] X. Zheng, C. Zhang, P. Chen, K. Zhao, H. Jiang, Z. Jiang, H. Pan, Z. Wang, and W. Jia. "A CRNN System for Sound Event Detection Based on Gastrointestinal Sound Dataset Collected by Wearable Auscultation Devices", *IEEE Access*. 2020.

[12] R. Rana. "Gated recurrent unit (GRU) for emotion classification from noisy speech", *arXiv preprint arXiv:1612.07778*. 2016.

[13] N. Minh-Tuan and Y. Kim, "Bidirectional Long Short-Term Memory Neural Networks for Linear Sum Assignment Problems", *Applied Sciences*, vol. 9, no. 17, p. 3470, 2019.

[14] C. Chen and Q. Li, "A Multimodal Music Emotion Classification Method Based on Multifeature Combined Network Classifier", *Mathematical Problems in Engineering*, vol. 2020, pp. 1-11, 2020.

[15] J. Li, X. Zhang, M. Sun, X. Zou and C. Zheng, "Attention-Based LSTM Algorithm for Audio Replay Detection in Noisy Environments", *Applied Sciences*, vol. 9, no. 8, p. 1539, 2019.

[16] Z. Zhang, S. Xu, T. Qiao, S. Zhang and S. Cao, "Attention Based Convolutional Recurrent Neural Network for Environmental Sound Classification", *Neurocomputing*, 2020.

[17] K.J. Piczak, "ESC: Dataset for Environmental Sound Classification", In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018). 2015.

[18] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. "librosa: Audio and music signal analysis in python". In *Proceedings of the 14th python in science conference*. 2015.

[19] M. Heideman, D. Johnson and C. Burrus, "Gauss and the history of the fast Fourier transform". *IEEE ASSP Magazine*, 1(4), pp.14-21, 1984

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. J*ournal of Machine Learning Research*, *15*(1), pp.1929- 1958. 2014.

[21] The top 10 causes of death, WHO, 2021. [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Accessed on 5 Jan].

[22] A. Marques, A. Oliveira and C. Jácome. "Computerized adventitious respiratory sounds as outcome measures for respiratory therapy: a systematic review". *Respir Care* 59(5):765–776. 2014.

[23] Y. Tang, J. Schmitz, D. Persing and C. Stratton, "Laboratory Diagnosis of COVID-19: Current Issues and Challenges", *Journal of Clinical Microbiology*, vol. 58, no. 6, 2020.

[24] X. Shen, Q. Dai, F.L. Chung, W. Lu and K.S. Choi. "Adversarial Deep Network Embedding for Cross-Network Node Classification". In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2991-2999). 2020.

[25] S.C. Neoh, W. Srisukkham, L. Zhang, S. Todryk, B. Greystoke, C.P. Lim, A. Hossain and N. Aslam. "An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images," *Scientific Reports*, 5 (14938). 2015.

[26] W. Srisukkham, L. Zhang, S.C. Neoh, S. Todryk and C.P. Lim. "Intelligent Leukaemia Diagnosis with Bare-Bones PSO based Feature Optimization," *Applied Soft Computing*, 56. pp. 405-419. 2017.

[27] L. Zhang, K. Mistry, S.C. Neoh. and C.P. Lim. "Intelligent facial emotion recognition using moth-firefly optimization," *Knowledge-Based Systems*. Volume 111, Nov. 2016, 248–267.

[28] T. Tan, L. Zhang and C.P. Lim. "Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks," *Knowledge-Based Systems*. 2019.

[29] T. Tan, L. Zhang, C.P. Lim. B. Fielding, Y. Yu, and E. Anderson. "Evolving Ensemble Models for Image Segmentation Using Enhanced Particle Swarm Optimization," *IEEE Access*. 2019.

[30] L. Zhang, C.P. Lim and Y. Yu. "Intelligent human action recognition using an ensemble model of evolving deep networks with swarm-based optimization". *Knowledge-Based Systems*, 220, p.106918. 2021.

[31] D. Farid, L. Zhang, C.M. Rahman, A.M. Hossain and R. Strachan. "Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-class Classification Tasks," *Expert Systems with Applications*. Volume 41, Issue 4, Part 2, March 2014, 1937–1946. 2014.

[32] P. Kinghorn, L. Zhang and L. Shao. "A Hierarchical and Regional Deep Learning Architecture for Image Description Generation," *Pattern Recognition Letters*. 2019.

[33] L. Zhang and C.P. Lim. "Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models", *Applied Soft Computing*. 92, 106328. 2020.

[34] Z. Zhu, X. Fan, X. Chu and J. Bi. "HGCN: A Heterogeneous Graph Convolutional Network-Based Deep Learning Model Toward Collective Classification". In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1161-1171). 2020.

[35] T. Tan, L. Zhang and C.P. Lim. "Intelligent skin cancer diagnosis using improved particle swarm optimization and deep learning models," *Applied Soft Computing*, p.105725. 2019.

[36] K. Mistry, L. Zhang, S.C. Neoh, C.P. Lim, and B. Fielding. "A micro- GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition," *IEEE Transactions on Cybernetics*. 47 (6) 1496– 1509. 2017.

[37] Y. Zhang, L. Zhang, S.C. Neoh, K. Mistry and A. Hossain. "Intelligent affect regression for bodily expressions using hybrid particle swarm optimization and adaptive ensembles," *Expert Systems with Applications*, 42 (22). pp. 8678-8697. 2015.

[38] P. Kinghorn, L. Zhang and L. Shao. "A region-based image caption generator with refined descriptions," *Neurocomputing*. 272 (2018) 416-424.

[39] T. Lawrence, L. Zhang, C.P. Lim and E. Phillips. "Particle Swarm Optimization for Automatically Evolving Convolutional Neural Networks for Image Classification", *IEEE Access.* 2021.

[40] H. Xie, L. Zhang and C.P. Lim. "Evolving CNN-LSTM Models for Time Series Prediction Using Enhanced Grey Wolf Optimizer", *IEEE Access*. 8, p. 161519-161541. 2020.

[41] B. Fielding and L. Zhang. "Evolving Image Classification Architectures with Enhanced Particle Swarm Optimisation," *IEEE Access*. 2018.

[42] S.C. Neoh, L. Zhang, K. Mistry, M.A. Hossain, C.P. Lim, N. Aslam and P. Kinghorn. Intelligent Facial Emotion Recognition Using a Layered Encoding Cascade Optimization Model. *Applied Soft Computing*. Volume 34, 2015, 72–93. 2015.

[43] D. Pandit, L. Zhang, S. Chattopadhyay, C.P. Lim, and C. Liu. "A Scattering and Repulsive Swarm Intelligence Algorithm for Solving Global Optimization Problems," *Knowledge-Based Systems*. 2018.

[44] M.M. Ghazi, B. Yanikoglu and E. Aptoula. "Plant identification using deep neural networks via optimization of transfer learning parameters". *Neurocomputing*, 235, pp.228-235. 2017.

[45] P. Kinghorn, L. Zhang and L. Shao. "Deep learning based image description generation", In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 919-926, 2017.

[46] B. Fielding and L. Zhang. "Evolving Deep DenseBlock Architecture Ensembles for Image Classification". *Electronics*. Nov 2020.

[47] Y. Yu, X. Chen, L. Zhang, R. Gao and H. Gao. "Neural Graph for Personalized Tag Recommendation". *IEEE Intelligent Systems*. 2020.

[48] S. Slade and L. Zhang. "Topological Evolution of Spiking Neural Networks". In *Proceedings of IEEE World Congress on Computational Intelligence*. IEEE, Brazil, 2018.

[49] B. Fielding, T. Lawrence and L. Zhang. "Evolving and Ensembling Deep CNN Architectures for Image Classification". In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*. 2019.