A Self-Attention Based Hybrid CNN-LSTM Architecture for Respiratory Sound Classification

Paritosh Bhushan[1], Md. Shah Fahad[2,*], Sanjay Agrawal[1], Kota Sai Durga Kamesh[1], Garima Singh[3], Paritosh Tripathi[1], Praveen Mishra[1], Vineet Kumar Singh[1], and Akshay Deepak[1]

ARTICLE INFO

Article history:
Received: 21 July 2022
Revised: 15 September 2022
Accepted: 4 October 2022

Keywords:
CNN-LSTM
Crakles
Respiratory sounds
Self-attention
Wheezels

ABSTRACT

For automating the diagnosis of respiratory and pulmonary diseases identification of breathing anomalies (wheeze, crackle) have played vital role. In this respect, recent application of machine learning techniques for sound anomaly classification have delivered promising results. This analysis can lead to a two class (healthy and diseased), three class (wheeze/crackle/healthy) or four class (wheeze/crackle/both/healthy) problems. Most of the existing works deal with two or three class problems. Among the few dealing with the four class problem, patient-dependent setup has been considered. Patient-independent modelling is essential to generalize the results for real-life scenarios. We could find only one work with a patient- independent setup with a reported accuracy of 49.5%. Here, we propose a deep CNN-LSTM + Self- Attention model that treats the above as a four class problem and is modelled towards being patient-independent. On the ICBHI'17 database [37], our proposed model achieves an improvement of 7.62% over state-of-the-art. The improvement is the result of the use of self-attention on LSTM networks along with data normalization.

1. INTRODUCTION

As of 2019, around 7% of all non-communicable disease (NCD) deaths take place due to chronic respiratory diseases (CRDs), which are largely avoidable via public policies that tackle their common threat factors [1]. In this regard, accurate lung auscultation is critical for disease diagnosis and assessment. Hence, there is a strongest interest in automating sound analysis of the lung. The digital respiratory sounds provide critical clinical characteristics such as regular and pathological indexes, which are essential for telemedicine and smart diagnosis [2]. Wheeze and crackle are the two most medically important lung sound anomalies. A wheeze is a sharp and consistent sound that happens while a breathing airway is obstructed. Crackles are often linked to obstructive lung disorders, such as asthma.

The lungs are one of the most important organs in our bodies, yet they are frequently overlooked. Asthma, acute lower respiratory tract infections, tuberculosis, and lung cancer, Chronic obstructive pulmonary disease (COPD) are the leading factors of death and disability [3].Although there is considerable clinical interest in a computerised analysis of respiratory sounds, use of machine learning in this field is still in its formative stage. Since lung sounds are unchanged and non-linear signals, they are difficult to illustrate and segregate. Here, the use of an electronic stethoscope has facilitated automated analysis [34].

Several attempts have been made in the last decade to automate the detection of breathing irregularities; these efforts consist of better machine learning algorithms and feature extraction techniques. random forest [4],The Dynamic Time Warping (DTW), Gaussian mixture model (GMM) [9], Hidden Markov Model (HMM)[10] and logistic regression model (LRM)[8] are some of the previously used machine learning based methods for the classification of respiratory sounds to detect breathing sound anomalies. The task of identifying breathing irregularities can be modelled as a two-class problem (healthy/diseased) [13], three-class problem (wheezes/ crackles/ healthy) [10,4] or four-class problem (wheezes/crackles/healthy/both wheezes and crackles) [12, 15] [11].

To distinguish the breathing cycles into two classes, namely healthy and diseased, perna et. al. [13] employed a features (MFCC). They achieved a precision of 83% by utilizing an 80-20 train-test split on a patient- dependent setup. They also classified the recordings into three

[1]*Institute of Engineering and Technology, Dr. Ram Manohar Lohia Avadh University, Ayodhya U.P., India.*
[2]*Birla Institute of Technology Mesra, Ranchi, India*.
[3]*Department of CSE, National Institute of Technology, Patna, Bihar, India.*
**Corresponding author:* Phone: +91-9973766537; Email: fahad8siddiqui@bitmesra.ac.in.

categories: stable, chronic, and non-chronic diseases, with an accuracy of 82%.

On a smaller subset of the ICBHI'17 dataset[37] consisting of 489 recordings, Chen et. al. used optimised S-transform dependent feature maps and deep residual nets (ResNets) to categorise the samples (not particular breathing cycles) into three classes, namely healthy, crackles, and wheezes. In a patient-dependent setup on a 70-30 train-test split, their model achieved 98.79% accuracy.

Table 1. A Comparison of available literature On ICBHI 17 DATASET

References	Features	Classifiers	Results
2-Class			
Perna et al.\cite{b13}	MFCC	CNN	Accuracy: 83%(for the healthy-unhealthy classification in 2-Class, 80-20 split)
3-Class			
Jakovljevic et al.\cite{b10}	MFCC	GMM+HMM	Accuracy: 39.56% (original train test split), 49.5% (training data, using 10-fold cross-validation) Accuracy: 82% (3-class classification using healthy, chronic and non-chronic)
Chen et al.\cite{b14}	Optimized S-transform	ResNets	Accuracy: 98.79% (subset of original data (smaller), 70-30 split test for 3-class sample level classification)
4-Class			
Kochetov et al.\cite{b12}	MFCC	Noise masking RNN	Accuracy: 65.7% (80-20 split test for the 4-class classification)
Arindam et.al\cite{b11}	MFCC	Hybrid CNN+RNN	Accuracy: 66.31% (four-class problem), 71.81% (retraining the patient specific data)

For the four-class problem (wheezes /crackles/ healthy /both wheezes and crackles), Kochetov et al.[12] suggested a noise labelling Recurrent Neural Networks. There are two sections in proposed model: (i) an attention network proposed for binary respiratory cycle groupage into noisy and non-noisy groups and (ii) an RNN for the 4-class classification. The RNN receives the filtered audio after the attention network has learned to recognise noisy pieces of the sound and suppresses them. They achieved a score of 65.7% with an 80-20 split in a patient-dependent setup. All of the above experiments are in a patient-dependent setup.

The research works discussed next were modelled in a patient-independent manner. In a patient-independent setup, a non-overlapping set of patients are used for training and testing. That is, the model trains on a known set of patients but it is tested on an unseen set of patients – much like the way patients would be encountered in real life scenarios. Such a setup is challenging due to the variation in individual features of the patients.

To distinguish the breathing cycles into three groups, namely healthy, wheezes, and crackles, Jakovljevic et. al. [10] used a HMM. They worked on a 3-class problem with the setup as patient independent. They pre-processed the data using spectral subtraction-based noise suppression and then classified it using the MFCC features. On the initial train-test split, their models scored 39.56% and on a 10-fold cross-validation method of the training sample, they scored 49.5%.

G. Chambres et al.[15] introduced a patient-level model in which low-level features (melbands, MFCC etc.), rythm characteristics (loudness, bpm, etc.), SFX characterstics (harmonicity and inharmonicity information), and tonal features are used to categorise particular breathing cycles from one of the four classes (four classes chords strength, frequency, tuning, etc.). They employ boosted tree process for the classification. Based on the percentage of breathing cycles classification as irregular the patients were then categorised as healthy or unwell. They achieved accuracy of 49.63% a breathing cycle level classification and 85% accuracy of a patient level classification.

Arindam et al proposed CNN-RNN hybrid model for the ICBHI'17 medical challenge respiratory sound database [37] on four-class classification of breathing cycles that achieves a score of 66.31%. If the model is retrained by patient-particular information, it achieves a leave-one-out validation score of 71.81%. Here, the pre-training was done in a patient-dependent manner. This creates the data-leakage problem in the retrained model. That is, even though the retraining is done in a patient-independent setup.

From the above pretext, it is clear that only a few have worked on the original 4-class problem of respiratory sound classification with the best breathing cycle level classification accuracy of 49.63%. Convolutional neural networks (CNNs) have been successfully applied in a wide variety of fields: from image processing [35] to genomic sequencing [36]. CNNs and RNNs have been found to be an

effective method for classifying abnormal breath sounds in prior studies [11–13].

In this proposed work, we have dealt with the respiratory sound classification problem with 4 classes: wheezes, crackles, both (wheeze and crackle), and healthy. Some well-liked features as Spectrogram [4], MFCC [5], Wavelet Coefficients [6], entropy-based features [7], and other extraction techniques were used. Further, we have implemented LSTM+ CNN + Self-Attention machine learning algorithms on the original settings of the 4-class problem to achieve a greater accuracy score.

A main concern in preparing a deep network is that it needs a huge dataset as well as a substantial amount of time. The first problem is compounded in the medical science because of the field of medical datasets are scarce and complex to come by. One method to get around this problem is to employ data augmentation techniques like noise injection, different speed, random transferring, and pitch shift to make new synthetic data from the existing ones.

The highlighted research gap in the existing experiments done so far is that most of them concentrated on patient-dependent model. This dependency of results on patient information showed fairly good accuracy results but is disadvantageous particularly due to the variances seen between individuals and the large number of model parameters that are needed. The hybrid CNN- RNN model proposed by Arindam at al.[11] where a patient-specific model was developed using pre-trained model showed a considerably better accuracy score of 66.31%, but posed a major issue of data leakage by 70:30 ratio splits as it was not modelled to be patient independent. It also created separate patient specific models for each and every patient which posed an issue of increased usage of resources (time and money) and memory units. Therefore, most of the researchers have worked upon patient-dependent model thus achieving fairly good results and those who have worked upon bringing patient independence to the modelling, suffer from data leakage and patient specificity.

In this proposed work, we have used patient-independent data for training and testing our models to produce more generalised results for unseen data in the future. Two types of improvements are performed: model-wise improvement and achieving patient independent model. We have worked in this field by dividing our dataset into various folds and then working on them such that training a model on one fold of the patient data and then testing it on another fold of the patient data to achieve patient independence. It also produces comparatively better outcomes for unpredictable input values as training and testing data have been kept separate to achieve patient-independence. We have also worked on exploring deep neural networks and applied various hybrid model namely Simple Attention mechanism with Hybrid CNN-LSTM, Self- Attention mechanism with Hybrid CNN-LSTM, Attention mechanism with Hybrid CNN-GRU and Transformers.

Based on the above discussion, the contributions of the study of our work are as follows:

Patient Independent model:

We applied normalization and cross-validation approach for patient independent model for classifying breathing sounds into four categories on the scientific challenge respiratory sound of the ICBHI'17 data set [37] and obtained accuracy score of 54.28% and accuracy percentage of 53.56% which is extremely creditable as far as medical datasets are concerned. Methods like Long Short Term Memory Model, Self-Attention, Gated Recurrent Units (GRUs) have shown promising results in classification problems of other fields.

Further, we have explored recent advancements in deep learning like Simple attention, Self-attention, CNN-LSTM-Attention, Hybrid CNN-GRU-Attention on the dataset to achieve fairly good results and a better accuracy score than the base paper. We have combined CNN and LSTM networks with 2 types of attention: simple attention and self-attention to achieve better results. Attention process is an attempt of execution the same action of specifically focusing on a few relevant things, while overlooking others in deep neural networks. This is a technique for reformulating the term representation based on the learnt correlations with all terms in the sequence [16, 17] Our patient-independent model based on LSTM + CNN + Self Attention Masking produces an accuracy score of 57.02% and accuracy of 58.62%. On the other hand, our proposed CNN + GRU + Self -Attention produces an accuracy score of 54.54% and accuracy of 60.14% which consequently gives an improvement of +7.62% over state-of-the-art method used in base paper.

The sections are described as follows: The proposed model describes in section 2. Experiments and results discusses in section 3. And finally, section 4 concludes the paper.

2. PROPOSED MODEL

2.1 Feature Extraction by Spectrograms

In audio spectral analysis and other applications, the spectrogram is a fundamental tool. It has a long history in usage for speech processing. The spectrogram is an intensity plot of the Short-Time Fourier Transform (STFT) magnitude (usually on a log scale, such as dB). Parameters of the spectrogram include the:

1. window length M
2. window type (Hamming, Kaiser, etc.),
3. hop-size R
4. FFT length N

Music, linguistics, sonar, radar, speech recognition, seismology, and other fields use spectrograms extensively [24]. Audio spectrograms may be used to phonetically classify spoken words and analyse different animal calls [21, 22]. In this proposed work, we propose models based on

cutting-edge Neural network approaches that classify respiratory sounds using Mel-spectrogram.

We have converted speech signals to Mel-spectrogram for feature extraction by using librosa library and fed them as input to a CNN network that extracts abstract characteristics maps from them. Each breathing cycle is changed over completely to a 2-dimensional picture, with row corresponding to Mel scale frequencies and column corresponding to time (window), and every value representing the signal's log amplitude value for that frequency and time window. Since the sampling frequencies of the audio samples in the dataset differed, all the signals were first down sampled to 4kHz. Down sampling the sound samples to 4kHz outcomes in no deficiency of significant information since crackle and wheeze both signals are regularly available inside the recurrence range from 0 to 2kHz. After that, a Mel-frequency spectrum with a 240-ms window size and a 50% overlap with a hop length (number of samples between successive frames) of 120 is used.

2.2 Data Augmentation

Data augmentation refers to techniques for expanding how muchinformation is by inserting slightly changed copies of existing information or making new artificial data from existing data. When a machine learning model trained, it serves as a regularize and helps to minimise overfitting.

2.3Materials and Methodology

We have used a sequential network consisting of CNN, LSTM and Completely linked (Fully Connected) softmax layers. CNN assign importance (learnable weights and biases) to various aspects/objects in the input images so that one can differentiate from other [26]. Thus, pre-processing is reduced significantly. Due to the intermittent character of both wheeze and crackle, as well as transient and frequency variety, mixture CNN-RNN structures may be valuable for lung sound grouping.

This is critical when designing an architecture that is capable of learning features while still being scalable to large datasets. It is noted that ConvNet's give good results in the audio classification with the help of passing log MEL Spectrograms [20].

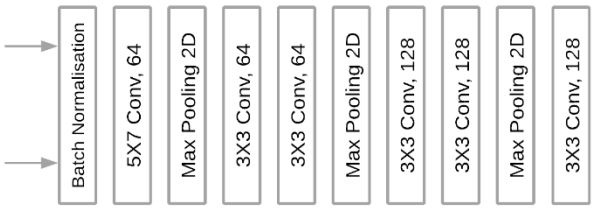

Fig. 1. CNN Architecture for Proposed Model.

Before passing the padded log MEL Spectrograms into the Model, it is necessary to pass the spectrograms to the masking layer, to filter out un timestamped data. To extract features based on timestamps, we used a four-stage model. The first stage consisting of a deep CNN model that extracts distinct characteristics of the given data. The 2- stage consisting a recurrent neural network layer that learns temporal relations [23]. The third stage consists of attention module that identifies the important features for better classification [25]. The fourth stage consists of a dropout layers, completely linked layers and softmax layers that translate the output data. The first stage is made up of the Batch-normalization, 2D stacked convolution, and max-pooling layers. For each mini-batch, batch normalisation standardises the inputs to a sheet. As a result, the learning process becomes more stable [27], and the number of training epochs required for training deep neural networks is significantly reduced. The output from the batch normalisation layer is convolved with a series of two 2D kernels in the 2D stacked convolution network to produce abstract feature maps with the same padding, resulting in the same output shape. The Rectified Linear activation functions follow each convolution layer which can reduce the likelihood of gradients to vanish [28]. The max-pool layer chooses the greatest values from a neighbourhood pixel, resulting in a reduction in total network parameters and shift-invariance. The same process is repeated for 3 times as shown in the Figure 1 to extract the features with high-level information from the spectrograms.

The second stage consists of LSTM networks, which were proposed by Sepp Hochreiter and Ju¨rgen Schmidhuber in 1995 [29] is an RNN that can learn order dependency in sequence prediction problems. A standard LSTM unit consists of a cell which remembers values over arbitrary time intervals, is regulated by the three gates namely input gate, output gate, and forget gate. To learn the temporal features, we used the non-linear activation function *tanh*.

The third stage is the attention module where we implement the activity of specifically focusing on a few of important things [25], while ignoring others in deep neural networks in which we tried different approaches which include the following:

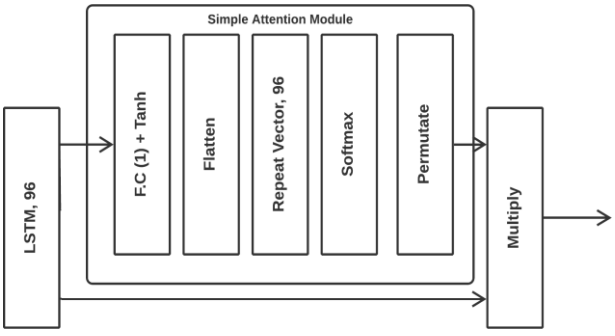

Fig. 2. Simple Attention Network.

LSTM with Simple Attention: The features we obtained from the LSTM require some attention in order to improve classification accuracy. To begin, the obtained features are passed to the dense layer, where they are applied the softmax activation function and permuted to obtain the weights. These weights indicate the individual attention for each time step. We multiply the focus weights with the original features to give certain features more importance while ignoring others [30]. The simple attention module which was described is shown in Figure 2.

With a sequence of high level representations, an attention layer is used to focus on emotion compitable parts and produce discriminative utterance- level representations for speech emotion recognition. In this paper, we use an attention model to score the importance of a sequence of high-level representations to the final utterance-level emotion representations, instead of simply performing a mean/max pooling over time.

$$\alpha_t = \frac{\exp(W.h_t)}{\sum_{t-1}^{T} \exp(W.h_t)} \quad (1)$$

$$c = \sum_{t-1}^{T} \alpha_t h_t \quad (2)$$

Explicitly, as describe in Fig.2, with the LSTM output ht=[h_t,h_t] at time *t*, firstly we compute the normalized importance weight α_t by a softmax function as in (1). After that we calculate the utterance-level presented *c* by performing a weighted sum on according to the weights (2) [38].

LSTM with Self Attention*:* When we think about the English word "attention" we know that it refers to focusing your attention on something and paying closer attention. Deep Learning's Attention mechanism is built on the idea of directing the focus, and it pays more attention to certain aspects when processing the data. This is a technique for reformulating the term representation based on the learned correlations with all terms in the sequential data. Generally, two types of attention mechanisms are used, *General Attention* and *Self Attention*. In *General Attention*, the interdependence is calculated between input and output elements, whereas in *Self Attention* the interdependence is calculated within the input elements.

In SER, we have a sequential input in the form of speech signal and a one-word output representing the emotion class. Since it is a classification-based problem, we do not have any output sequence. And as mentioned above the *General Attention* mechanism uses input sequence as well as output sequence for calculating interdependence. But here the output sequence BiLSTM is just one word which may not be able to calculate as much significant attention as a sequence does. So we decided to use *Self Attention* in place of *General Attention* which calculates attention within the input sequence only.

For calculating self-attention, we have to create three vectors from each of the encoder's input vectors. So for each word (here word is used for the vector), representing a small segment of the input speech segment of 3 seconds, we create a Query vector, a Key vector, and a Value vector. These vectors are created by multiplying the embedding by three matrices that we trained during the training process.

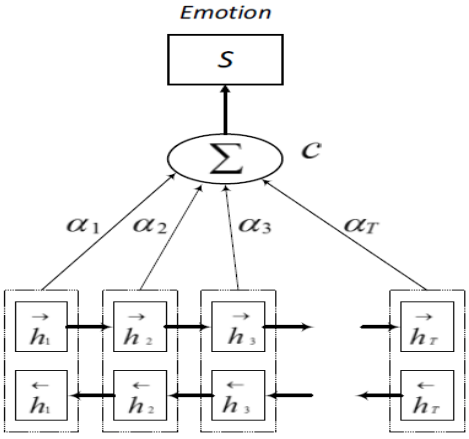

Fig.3. The attention layer working process

$$Qi = W_q xi \quad (3)$$

$$Ki = W_k xi \quad (4)$$

$$Vi = W_v xi \quad (5)$$

From these three vectors, the self-attention matrices are calculated as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Since not all structure level CNN-BiLSTM characteristics contribute uniformly to the representation of the speech emotion, a Self-Attention layer has been used with a series of supreme-level representatives to concentrate on emotion pertinent parts and generate discriminative utterance-level representations for SER. In the place of simply conducting a mean/max pooling over time, we utilize the Self-Attention model to score the significance or a series of top-level representations to the last utterance-level emotion delineation. We calculate Self-Attention by passing it from a Self-Attention layer defined in *keras self-attention* with activation as 'relu' and the attention width as 15 and then perform *Global-Average-Pooling*. Self-attention is considered as a separate layer that is coupled with the CNN and LSTM models to more fully integrate their respective strengths.

The fourth stage consists of dropout, fully linked (FC) and softmax layers. Here, the dropout layer sets input units to 0 with a frequency of rate at each phase during training period preventing the model from overfitting [31]. The output features from the Attention Module are passed to the dropout Layer, where they are further processed to the fully connected dense layer, where Relu is used as the activation function. Then it's passed back to the dropout layer, which then applies it to the softmax activation function, which calculates a likelihood for each possible class.

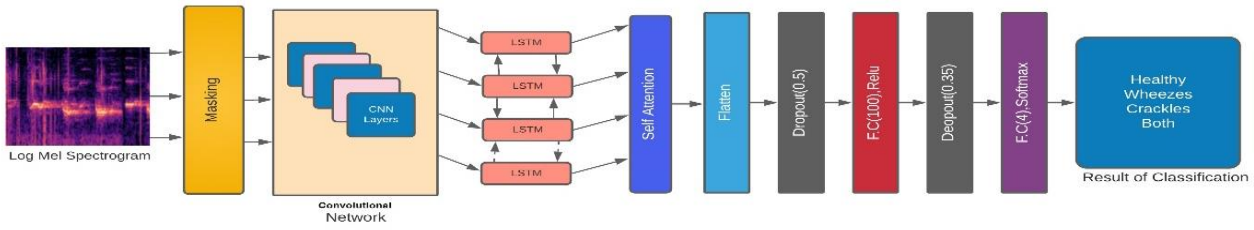

Fig. 4. Comparison of the proposed architecture with other deep learning based architecture.

3. EXPERIMENTS AND RESULTS

3.1 Dataset description

We used the scientific challenge respiratory sound database from the ICBHI'17[37] which consists of 920 annotated audio samples from 126 subjects.

The database contains a total of 5.5 hours of recordings containing 6898 respiratory cycles, of which 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes. Data is collected from various positions on the chest viz. trachea, left and right anterior, left and right posterior, left and right lateral.

The four columns in the annotation files are starting of respiratory cycle(s), end of respiratory cycle(s), presence/absence of crackles (presence denoted as 1, absence denoted as 0), presence / absence of wheezes (presence denoted as 1, absence denoted as 0). LRTI stands for Lower Respiratory Tract Infection, and URTI stands for Upper Respiratory Tract Infection in the diagnosis file. Furthermore, a large percentage of the samples of data is found to be noisy, which simulate real life conditions.

3.2 Results

In this paper, we explored the effectiveness of using models that are based on self-attention with CNN. The proposed hybrid CNN-LSTM with Self Attention model achieves a score of 57.02% and accuracy of 58.62% on patient-independent data of the 4-class grouping of breathing cycle for ICBHI'17 database which is represented in Table 2. The comparison with different models and the proposed model is shown in the table 3. This proposed model gives significantly more authentic results as model based on self-attention with CNN outperform the rest. The improvement is the result of the use of self-attention on LSTM networks along with data normalization. Further, the proposed model is patient independent to prevent data leakage. On hybrid CNN-GRU with self-attention model and masking with patient-independent, the average score obtained is 54.54% and the accuracy obtained is 60.14%.

Here We have used self-attention on GRU network. Self-attention is also use to get the long range dependencies of feature. We have also worked on a hybrid CNN-LSTM with simple attention model that produces a score of 52.8% and accuracy of 53.96%. On performing 10-fold cross-validation on the dataset for Hybrid CNN-RNN model, the average score obtained is 68.61% in which the model also outperform results reported by Arindam et. al. [11].

Table 2. Confusion Matrix using Self-attention based hybrid CNN-LSTM architecture

Class	Wheezes	Crackles	Healthy	Both
Wheezes	**48.62**	6.42	33.03	11.93
Crackles	1.05	**57.54**	40.35	1.05
Healthy	2.75	29.80	**66.27**	1.18
Both (wheezes and crackles)	37.50	29.55	25.00	**7.95**

Table 3. Comparison of the proposed framework with other deep learning based architecture

Model	Score	Accuracy
Arindam et al. (Base Paper)	49.58	50.97
CNN+LSTM+Simple-Attention	52.80	53.96
CNN+GRU+Self-Attention	54.54	**60.14**
CNN+LSTM+Self-Attention	**57.02**	58.62

In the table 2, we are getting good score for healthy class and lowest for both. This is because there are some features that are colliding with both wheezes and crackles which results in less accuracy for the both classes.

4. CONCLUSION

In this paper, our main motive is to achieve patient independence in our model to produce generalised results for unseen data and to reduce data leakage and memory usage. We explored deep neural networks like simple attention, self -attention and Gated Recurrent Units to the Mel-spectrograms produced. Our model consists of 4 main layers: the initial masking layer, CNN-LSTM layers, self-attention layer and, lastly, dropout, fully linked (FC) and softmax layers with all

the important model architecture hyper parameters tuned appropriately to minimise training failure. Self- attention is considered as a separate layer that is coupled with the CNN and LSTM models to fully integrate their respective strengths. This introduced model gives significantly more dependable outcomes accomplishing a score of 57.02% for the original train-test split which results in a gain of +7.44% accuracy score than Arindam et. al model. The experimental results show that the proposed model of CNN+LSTM+Self-attention outperforms the hybrid CNN+LSTM, CNN+LSTM+Simple attention and CNN+GRU+Self-attention model in terms of the overall accuracy score. Finally, we have achieved to secure patient independence results in our model along with bringing model-wise improvements and reducing computational complexity and memory foot- print present in existing literature on the ICBHI dataset.

REFERENCES

[1] Khaltaev N, Axelrod S.,2019. Chronic respiratory diseases global mortality trends, treatment guidelines, life style modifications, and air pollution: preliminary analysis. J Thorac Dis.;11(6):2643-2655.

[2] Naqvi, S.Z.H.; Choudhry, 2022. M.A. An Automated System for Classification of Chronic Obstructive Pulmonary Disease and Pneumonia Patients Using Lung Sound Analysis.Sensors2020,20(22),6512.

[3] Forum of International Respiratory Societies 2017. The Global Impact of Respiratory Disease–Second Edition. Sheffield, European Respiratory Society.

[4] J. Acharya, A. Basu, and W. Ser. 2017. Feature extraction techniques for low-power ambulatory wheeze detection wearables, Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society: 4574–4577.

[5] B.-S. Lin and B.-S. Lin. 2016. Automatic wheezing detection using speech recognition technique, Journal of Medical and Biological Engineering, vol.36, no.4: 545–554.

[6] M. Bahoura, 2009. Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes, Computers in Biology and Medicine, vol. 39, no. 9,: 824–843.

[7] J. Zhang, W. Ser, J. Yu, and T. Zhang, 2009. A novel wheeze detection method for wear- able monitoring systems, International Symposium on Intelligent Ubiquitous Computing and Education IEEE.Chengdu, China 15-16 May: 331–334.

[8] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux, 2016. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population, Computers in Biology and Medicine, vol.70: 40–50.

[9] I. Sen, M. Saraclar, and Y. P. Kahya, 2015. A comparison of svm and gmm based classifier configurations for diagnostic classification of pulmonary sounds, IEEE Transactions on Biomedical Engineering, vol. 62, no. 7,: 1768–1776.

[10] N. Jakovljevic and T. Lon car Turukalo, 2018. Hidden markov model based respiratory sound classification, in Precision Medicine Powered by pHealth and Connected Health. Springer: 39–43.

[11] J. Acharya and A. Basu, 2020. Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning, IEEE transactions on biomedical circuits and systems. Vol. 14: 535–544.

[12] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, 2018. Noise masking recurrent neural network for respiratory sound classification, International Conference on Artificial Neural Networks. Springer: 208–217.

[13] D. Perna, 2018. Convolutional neural networks learning from respiratory data, IEEE International Conference on Bioinformatics and Biomedicine (BIBM).: 2109–2113.

[14] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, 2019. Triple-classification of respiratory sounds using optimized s-transform and deep residual networks,"IEEE Access, vol. 7, 32 845–32 852.

[15] G. Chambres, P. Hanna, and M. Desainte-Catherine, 2018.Automatic detection of patient with respiratory diseases using lung sound analysis, International Conference on Content-Based Multimedia Indexing (CBMI). IEEE: 1–6.

[16] D. Bahdanau, K. Cho and Y. Bengio, 2015. Neural machine translation by jointly learning to align and translate, ICLR.

[17] H. Lamba, 2019. Intuitive Understanding of Attention Mechanism in Deep Learning, Towards Data Science

[18] Hou, N., Xu, C., Chng, E.S. and Li, H., 2019. Domain adversarial training for speech enhancement, Proceedings of 2019Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China 18-21 November: 667-672.

[19] A. Tripathi, A. Mohan, S. Anand and M. Singh, 2018. Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Calgary, AB, Canad April 15 - 20: 5959-5963.

[20] A. Meghanani, A. C. S. and A. G. Ramakrishnan, 2021. An Exploration of Log-Mel Spectrogram and MFCC Features for Alzheimer's Dementia Recognition from Spontaneous Speech, IEEE Spoken Language Technology workshop,: 670-677.

[21] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen. 2017. Convolutional recurrent neural networks for polyphonic sound event detection, IEEE/ ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6: 1291–1303.

[22] J. Sang, S. Park, and J. Lee, 2018. Convolutional recurrent neural networks for urban sound classification using raw waveforms, 26th European Signal Processing Conference (EUSIPCO) 03-07 September: 2444– 2448.

[23] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlu¨ter, Shuoyiin Chang, Tara Sainath. 2019. Deep Learning for Audio Signal Processing, Journal of Selected Topics of Signal Processing, Vol. 13, No. 2: 206–219.

[24] Abhilasha, Preety Goswami, Prof. Makarand Velankar, 2013, Study paper for Timbre identification in Sound, International Journal of Engineering Research & Technology (IJERT) Volume 02, Issue 10

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.2017. Attention Is All You Need, revised 6 Dec 2017 from https://arxiv.org/abs/1706.03762v5.

[26] Keiron O'Shea, & Ryan Nash. (2015). An Introduction to Convolutional Neural Networks, revised 2 Dec 2015from https://arxiv.org/abs/1511.08458v2.

[27] Sergey Ioffe, & Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Net- work Training by Reducing Internal Covariate Shift revised 2 Mar 2015, from https://arxiv.org/abs/1502.03167v3.

[28] Abien Fred Agarap. 2019. Deep Learning using Rectified Linear Units (ReLU) revised 7 Feb 2019 from https://arxiv.org/abs/1803.08375v2.

[29] Hochreiter, Sepp & Schmidhuber, Ju¨rgen. 1997. Long Short-term Memory, Neural computation. Vol. 9 issue 8.:1735-1780.

[30] Haoye Lu, Haolong Zhang, & Amit Nayak. 2020. A Deep Neural Network for Audio Classification with a Classifier Attention Mechanism, 14 Jun 2020 from https://arxiv. org/abs/2006.09815v1.

[31] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, & Ruslan Salakhutdinov, 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(56), 1929-1958.

[32] Yang, Shu-wen & Liu, Andy & Lee, Hung-yi. 2020. Understanding Self-Attention of Self-Supervised Audio Transformers revised 10 Aug 2020 from https://arxiv. org/abs/2006.03265v2.

[33] Sneha Chaudhari and Gungor Polatkan and Rohan Ramanath and Varun Mithal (2019). An Attentive Survey of Attention Models revised 12 Jul 2021, V3 from https://arxiv. org/abs/1904.02874v3.

[34] Shuang Leng, Ru San Tan, Kevin Tshun Chuan Chai, Chao Wang, Dhanjoo Ghista, and Liang Zhong, 2015. The electronic stethoscope, BioMedical Engineering OnLine 14(1):66.

[35] Browne M., Ghidary S.S. (2003) Convolutional Neural Networks for Image Processing: An Application in Robot Vision. In: Gedeon T.D., Fung L.C.C. (eds) AI 2003:AI 2003. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg vol 2903,: 641–652.

[36] Hemalatha Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, Chandran Venkatesan, C. Suresh Gnana Dhas, "Analysis of DNA Se- quence Classification Using CNN and Hybrid Models", Computational and Mathematical Methods in Medicine, vol. 2021: 12 pages.

[37] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A.Oliveira, C. Jacome, A. Marques et al. 2018, A respiratory sound database for the development of automated classification, Precision Medicine Powered by pHealth and Connected Health. Springer Nature Singapore Pte Ltd : 33–37.

[38] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang,2018. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition , IEEE Signal Processing Letters , Volume: 25, Issue: 10 : 1440-1444.