Hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) Based Classification of Respiratory Infections

Kaustubh Prabhu
CSE Dept *of UConn*
kaustubh.prabhu@uconn.edu

Shubhangi
CSE Dept *of UConn*
shubhangi.shubhangi@uconn.ed

Abstract— **The respiratory sounds indicate the physiology of the respiratory tract and their impairments due to the infections. This digital data recorded on digital devices unfolds the possibility of using predictive analysis to automatically diagnose respiratory disorders. There has been a good amount of work to classify healthy versus unhealthy sounds. However, based on these sounds most of the earlier work only classifies the lungs as normal or infected. Using the same respiratory sounds, in this paper, a DNN-HMM model-based classification of multiple respiratory diseases have been accomplished and successfully tested. Further, the performance of the developed DNN-HMM model is compared against a baseline HMM model and the superiority of the former is established.**

Keywords—DNN-HMM, gaussian HMM, MFCC, respiratory sound classification

I. Introduction

The respiratory sounds indicate the physiology of the respiratory tract including the throat to the nodal structure of lungs and the various impairments caused due to various infections. Normal lung sounds contain healthy bronchophony and healthy vesicular sound, the whole respiratory cycle is smooth with no sticky noise or roughness. The adventitious sounds emitted in the breath due to inflammation, obstruction, changes within lung tissue, and the position of secretions within the lung. They are sticky and have a lot of crackles and wheezes. The sounds indicate the upper respiratory tract infections, lower respiratory tract infections, Chronic Obstructive Pulmonary Disease, Asthma, Bronchitis, Pneumonia. Generally, it is difficult for a junior physician to easily classify these sounds into various classes of diseases. It is important to have a primary diagnosis to start basic treatment until the biopsy or other test results come. These sounds are being recorded using different types of digital stethoscopes and being sent to experts for reviewal. This digital data unfolds the possibility of using predictive analysis to automatically diagnose respiratory disorders. Hence there has been some work in this field and a lot of people have tried to create models for this kind of data to understand normal vs sick sounds. But there has been hardly any work to classify the diseases. One of the widely used dataset for it is the Respiratory Sound Database also, known as the ICBHI Scientific Challenge database, was created by two research teams in Portugal and Greece, contains audio samples collected independently over several years. It has data from various patients having various diseases.

Hence, we define our primary problem statement as creating a Model to classify various disease types given in our dataset. Taking a glance at the data makes us understand that there are a lot of things we need to take care of to explicitly state our objective. A detailed problem statement for this work would be as follows: Classification of respiratory sounds into various disease types using a Hybrid DNN-HMM model and compare it to a baseline model of a Gaussian HMM. Creation of which requires cleaning of data and representing it in a standardized metric form rather than audio, to feed into the model.

One of the main motivations for this work is to streamline the process of creation of primary diagnosis from respiratory sounds to give physicians and nurses a better understanding of patient's health using an unbiased state-of-the-art model. Also, to detect the respiratory diseases in no time and at a very initial stage. Moreover, in these covid times, this model could be more helpful for the patients and the doctors to conclude about lungs without spending much time on one patient diagnosis. Keeping it as a probabilistic model makes it better interpretable in case a doctor needs to be explained how the diagnosis was made.

II. Backgrounds and Related Work

A. Background

The World Health Organization identifies five respiratory diseases among the most common causes of severe illness and death worldwide, namely chronic obstructive pulmonary disease (COPD), asthma, acute lower respiratory tract infection (LRTI), tuberculosis, and lung cancer [1]. The number of people affected by COPD reaches 65 million, with about 3 million deaths per year, making it the third leading cause of death worldwide [2,3]. Asthma is a common chronic disease that is estimated in Jul 2019 to affect as many as 339 million people worldwide [4], and it is considered the most common chronic childhood disease. Another widespread disease that especially affects children under 5 years old is pneumonia [5]. The Mycobacterium tuberculosis agent has infected over 10 million people, and it is considered the most common lethal infectious disease [6]. Yet, lung cancers kill around 1.6 million people every year [7]. Prevention, early diagnosis, and treatment are key factors to limit the spread of such diseases and their negative impact on the length and quality of life. Lung auscultation is an essential part of the respiratory examination and helps diagnose various disorders, such as anomalies that may occur in the form of abnormal sounds (e.g., crackles and wheezes) in the respiratory cycle. When performed through advanced computational methods, a deep analysis of such sounds can be of great support to the physician, which could result in enhanced detection of respiratory diseases.

B. *Related Work*

Stethoscopes are applied as one of the most popular methods to diagnose pulmonary diseases. Compared with modern advanced medical treatments such as blood tests or X-ray chest radiographs, auscultation using the stethoscope is more convenient and cost-effective as a non-invasive diagnostic method to detect respiratory diseases. Although the traditional diagnosis method is widely used, it is a subjective process that relies on a physician's experience and ability to discriminate various lung sound patterns. What's more, the human ear is not sensitive to low frequencies [8]. The experiments have been performed on feature extraction techniques and noted that MFCC is one of the best respiratory feature extraction methodology. MFCC can be considered as one of the most prominent features for the classification of lung sounds pertaining to pulmonary crackle and pleural friction rub classes [9]. In the field of speech/sound recognition, several classification approaches have already been explored. Recently, many researchers focused on the classification of lung sounds using a machine learning strategy, which consists of two common phases, feature extraction, and pattern classification. In recent years, a novel hybrid model architecture, Deep Neural Network - Hidden Markov Model (DNN-HMM), has been proposed and widely used in speech recognition.[10]. A deep neural network (DNN), which can capture the underlying nonlinear relationship among data, is the conventional multi-layer perceptron’s with many layers, where training is typically initialized by a pre-training algorithm [11].]. The ICBHI for respiratory sound analysis has been used earlier for classifying a different kind of diseases like COPD, LRTI, etc. using MFCC feature extraction along with recurrent neural networks (RNNs) neural network architecture [12]

III. Data

The Data is the Respiratory Sound Database also, known as the ICBHI Scientific Challenge database. It was created by two research teams in Portugal and Greece, working independently over several years collecting data for various studies. This data was all clubbed together for this competition. It contains audio samples collected independently over several years. All the recordings followed the computerized respiratory sounds analysis guidelines for short-term acquisitions. Which are the basic standards under which data is acquired and stored. The collection of respiratory sounds was from seven chest locations: trachea; left and right of anterior, posterior, and lateral lungs. Since it was all scrambled and from multiple studies, the number of samples per location varies hugely. The samples were collected in clinical and non-clinical (home) settings and on multiple different recording equipments. The data was essentially collected using one of the following tools AKG C417L Microphone (AKGC417L), 3M Littmann Classic II SE Stethoscope (LittC2SE), 3M Littmann 3200 Electronic Stethoscope (Litt3200), WelchAllyn Meditron Master Elite Electronic Stethoscope (Meditron). We have the details of the recording equipment along with the mode at which it was operated and also the location from which it was taken. All this data is represented in the file name.

The acquisition of respiratory sounds was performed on subjects of all ages, from infants to adults and elderly people. Subjects included patients with lower respiratory tract infections, upper respiratory tract infections, COPD, Bronchiolitis, and Bronchiectasis. The database consists of a total of 5.5 hours of recordings containing 6898 respiratory cycles.

The structure of the overall dataset is a base folder that has all the metadata and then a folder containing the audio files. The metadata consists of 3 files namely, a text file listing the diagnosis for each patient, a text file explaining the file naming format, a text file containing demographic information for each patient. The subfolder with the actual data consists of 920 ‘.***wav***’ sound files and their annotation ‘.***txt***’ files containing the start and end time of the respiratory cycle. A glimpse of one of the annotation files can be seen below in Table 1.

Table 1. The annotation file states that the respiratory cycles can last anywhere from 0.5 sec to 2.5 sec.

Start	End	Crackle	Wheeze
0.036	0.579	0	0
0.579	2.45	0	0
2.45	3.893	0	0
3.893	5.793	0	0
5.793	7.521	0	0
7.521	9.279	0	0

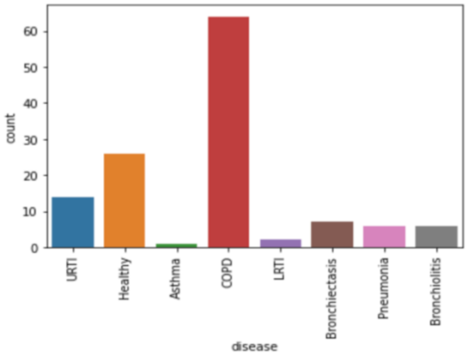


Fig. 1. Patient distribution per category.

A. *Data quality and Data Distribution*

As shown in Fig. 1, the data as per class distribution is highly imbalanced, out of the 126 patients more than half of them had COPD and two classes as seen in the fig below had a very low number of cases. The data has a lot of noise and that had to be handled. Since it was recorded in various conditions and different machines were used the noise was highly variable.

IV. Methods

We are proposing a framework in which we are using the best out of both the worlds, probabilistic and neural network. We are combining the hidden Markov model from a probabilistic approach and to enhance the limitations of HMM we are combining with DNN to form the DNN-HMM Model. To make good use of the realistic recordings, noise control is applied to reduce the interference of surrounding noises. We remove the extreme data outliers as inherently the respiratory data does not have high frequencies. We are also passing it through a wiener filter to smoothen it. Then the enhanced lung sounds are be broken down into actual respiratory cycles. Then MFCC is utilized to extract features and we are feeding the extracted features into the HMM layer and the output of HMM working as the Input of the DNN Model. We are also building the Gaussian-HMM with the same ICBHI data and comparing the DNN-HMM results with gaussian HMM. The flowchart of the proposed system is shown in Fig. 2.

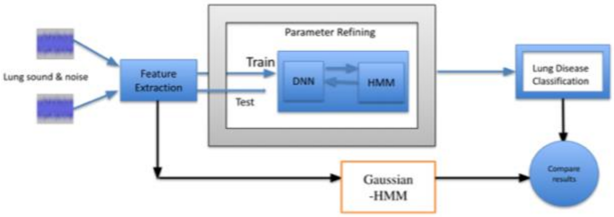


Fig. 2. System flow chart.

A. PreProcessing

The preprocessing includes separating each respiratory cycle, handle white noise, handle extreme noise and smoothen data. Firstly, with the help of annotation files for each audio file, we use its annotation file to break the larger file into smaller sections of 3-second length where it is padded with blanks to normalize data size. Then for each respiratory cycle, we remove white noise, then we use the trimmer to remove outliers. Then using the SCIPY Signal Library we run the data through a wiener filter which with the help of a few algorithms removes the constant variation to stabilize and smoothen the signal.

B. MFCC

The above-preprocessed data is passed through MFCC. It is one of the most widely used tools for feature extraction of sound/signal data. MFCC is especially beneficial when we are looking for a mixture of static and time-based features. The basic flowchart for MFCC is shown in Fig. 3. The idea over here is to do feature extraction of respiratory sound frequency by using MFCCs to describe the temporal characteristic of lung disease classification. The frequency bands are equally spaced on the mel-scale, which approximates the sound more closely than the linearly spaced frequency bands used in the normal cepstrum. This frequency warping allows for better representation of sound. A sample MFCC plot is provided in Fig. 4.

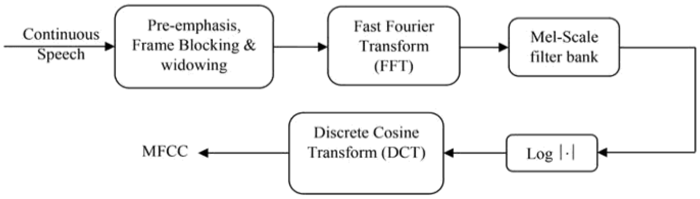

Continuous Speech
Pre-emphasis, Frame Blocking & widowing
Fast Fourier Transform (FFT)
Mel-Scale filter bank
Log |·|
Discrete Cosine Transform (DCT)
MFCC

Fig. 3. Flowchart of MFCC.

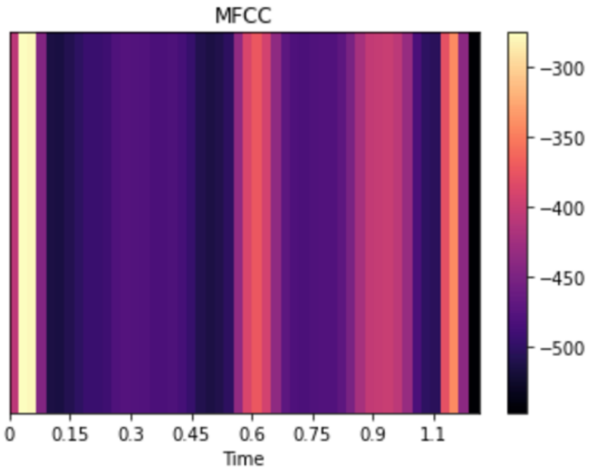


Fig. 4. MFCC plot.

C. Gaussian-HMM

HMM is a statistical Markov model with hidden states. When the data is continuous, each hidden state is modeled as Gaussian distribution.

The Gaussian hidden Markov model (Gaussian HMM) is a type of finite-state-space and homogeneous HMM where the observation probability distribution is the normal distribution

$$Y_t \,|S_t \sim N(\mu_{s_t}, \sum s_t) \tag{1}$$

where μ_{s_t} *and* $\sum s_t$ are mean and covariance parameters at state $S_t, S_t = 1, \dots., K$. Hence, the initial state probability vector (ISPV) π, the transition probability matrix (TPM)**A**, and the observation parameter **B** = (= $\{\mu_i, \sum_i\} i = 1 \dots k$, which consists of mean and covariance parameters) together specify the Gaussian HMM; that is, the parameter θ of the Gaussian HMM is $\{\pi, \boldsymbol{A}, \boldsymbol{B}\}$.

D. DNN:

A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs. Each hidden unit uses a nonlinear function to map the feature input from the layer below to the current unit. In our work, we use the traditional logistic function as the mapping function.

$$y = \frac{1}{1+e^{-(b+xw)}} \tag{2}$$

where x denotes the input feature, w denotes the weights between connections, b denotes the bias and y denotes the output unit.

DNN is capable of modeling very complex and highly nonlinear relationships between inputs and outputs, due to its flexible structure with multiple hidden layers and multiple hidden units.

DNN is capable of modeling very complex and highly nonlinear relationships between inputs and outputs, due to its flexible structure with multiple hidden layers and multiple hidden units. It's trained to classify the data gives out the class probabilities for each data point. That implies that the posterior/observation probabilities are inferred from the output of the DNN instead of the standard HMM taking input directly in form of MFCC.

Discriminative pre-training: As supervised pre-training, we followed. The proposed by referred to as discriminative pre-training (DPT). The general architecture. It works as follows, in a first step, a layer-wise Back Propagation (BP) is used to train a one-hidden-layer DNN to full convergence using every frame's state label, then the softmax layer is replaced by another randomly initialized hidden layer and a new random softmax layer on top, and the network is discriminatively trained again to full convergence. The process is repeated until the desired number of hidden layers is reached. As stated this is very similar to a greedy layer-wise training but differs in that of only by the updates of newly added hidden layers. showed that DNN supervised outperformed Gaussian HMM with pre-training.

While this achieved accuracies better than Gaussian HMM pretraining, we found that it can be further improved by stopping very early by going through the data only once and using a large learning rate The goal is to bring the weights close to a good local optimum.

E. DNN-HMM:

In conventional HMMs based sound recognition, the observation probabilities are modeled using Gaussian distribution under the maximum likelihood criterion. The potential of such a model is restricted since Gaussian are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space. To overcome this restriction, we propose a hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) for respiratory sound detection, where the output of the DNN is fed to the HMM.

Our DNN Model has Maxout Activation, with a generalization of ReLU, "learnable" activation function, while the cost function is managed using cross-entropy and for optimization, we use Momentum criterion with 0.95 and dropout was managed at 0.2 ratios. The stopping potential at 1.05. The model was validated with 10% data and a batch size of 240, for ten epochs, consisting of a sliding window of 14 units. The structure of the network had 256 base neurons and just two hidden layers with a scaling factor of 33%. The Output of which was then calculated by a final softmax layer. The HMM part of the model uses the observational probabilities to identify the transitional probabilities in the states and hence classify the diseases by giving out probable likelihoods, the following part goes into details for that

I. HIDDEN MARKOV MODEL:

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM represented as λ=(T,G,π), consists of the following elements:

1. The number of states in the model denoted as P, the set of states denoted as S={s1,s2,…, s_p }, and pt the state at time t.
2. T = $\{t_{ij}\}$ the state transition probability distribution with $t_{ij} = P(p_{t+1} = s_j | p_t = s_i)$;1≤ $i, j \leq P$
3. G = $\{g_i(o_t)\}$ the observation probability, where $g_i(o_t)$ represents the probability of observing o_t. G represents the finite mixture at state s_i GMM is a probabilistic model which can model N sub-population normally distributed. Each component in GMM is a Gaussian distribution.

GMM equation:

$$g_i(o_t) = \sum_{m=1}^{M} c_{im} \nsim (o_t, \mu_{im,} U_{im}), 1 \leq i \leq P \quad (3)$$

4. $\pi =$ an initial probability distribution over states. πi is the probability that the Markov chain will start in state i. Some states j may have πj = 0, meaning that they cannot be initial states. Also,

$$\pi_i = P(p_1 = s_i), 1 \leq i \leq Q \quad (4)$$

Now, to use HMM in DNN we need few things to be solved:

1. Learning problem: X = training set, learning procedure is to find the set of the model parameter $\leftthreetimes^* = \{T, G, \pi\}$ s.t $\leftthreetimes^*$=argmaxλP(X|λ), find the model parameters that better fit the training set. The forward-backward algorithm is used to calculate P(X|λ), while the Baum-Welch algorithm is employed to solve the learning problem
2. Decoding problem: Given a model λ and a sequence of new observations O=(o1,o2,…,oT) (referred to as testing set in the following), the decoding procedure is defined as the problem of finding the hidden state sequence (p1,…,pT) that have most likely produced that observation. The solution to this problem is given by the Viterbi algorithm.

In the case of respiratory sound detection, we train T HMMs {λt,(t=1,…,T} for T discrete emotion classes. For a new speech input O, it is assigned to the emotional class

II. DNN-HMM:

The key difference between DNN-HMM and GMM-HMM is the use of DNN (instead of GMM) to estimate the observation probabilities. We use the DNN to model p(p_t|o_t,), the posterior probability of the state given the observation vector o_t, which is possible since p(p_t) is easy to estimate from an initial state-level alignment of the training set.

1) DNN-HMM Training Procedure:

The detailed training process for emotion recognition is as follows:

1. For each emotion class T(t=1,…,T), left to right GMM-HMM λt with Q states is trained using the training speech sentences of class c.
2. For each speech sentence O=(o1,o2,…,oT) in the training set c, the Viterbi algorithm of the GMM-HMM, is performed on λc to obtain an optimal state sequence (qc1,…,qtT), and each state qct is assigned a label Li(i∈(1,…,T×Q)) according to a state-label mapping table.
3. All the training sentences, together with their labeled state sequences are used as inputs to train a DNN, whose outputs are the posterior probabilities of the T×P output units. pre-training, or (ii) the discriminative pre-training described below

2) DNN-HMM Recognition Procedure:

In the emotion recognition process, for an input speech sentence O=(o1,o2,…,oT), one should estimate the probability p(O|λc) for each emotion class c, and get the final recognition result. In GMM-HMM, this probability is obtained via the Viterbi algorithm.

In DNN-HMM, we adopt the following procedure to calculate the probability p(O|λt).

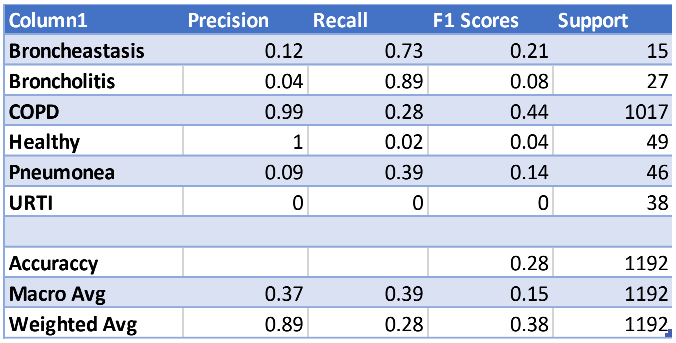
1. The input feature sequence is first input into the DNN, obtaining the posterior probabilities {p(Li|ot)}i=1,…,C×Q as outputs. Then the posterior probability p(qt=Sck|ot) can be obtained from p(Li|ot), by mapping the label Li to the state k of the model c, using a state-label mapping table.
2. According to the Bayesian principle, we calculate the likelihood probability p(ot|qt) as

In our implementation, the prior probability of each state, p(qt), is calculated from (occurrences of) the training set, and p(ot) can be assigned a constant since the observation feature vectors are regarded as independent of each other. For each emotion model λt, the Viterbi algorithm is performed to calculate the likelihood probability p(O|λt). However, here the probability bqt(ot) is replaced by p(ot|qt).

V. Results and Analysis

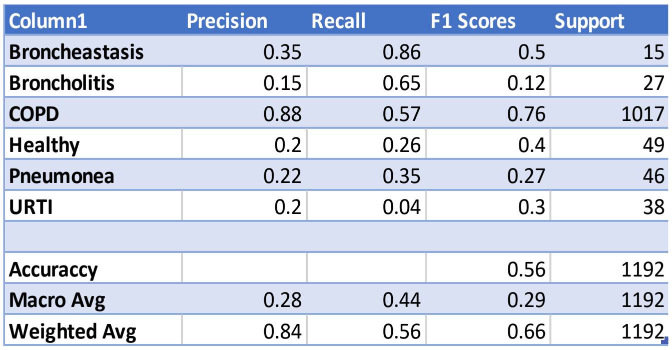
We have extracted respiratory sounds from ICBHI respiratory data by MFCC and build DNN-HMM and Gaussian model to classify the respiratory diseases into 6 classes including healthy. Moreover, we have compared the results in terms of precision, recall, F1score, and support(Precision : (true positive(tp)/ (tp + false positive(fp)) measures the ability of a classifier to identify only the correct instances for each class. Recall : (tp / (tp + fn) is the ability of a classifier to find all correct instances per class.F1 score: is a weighted harmonic mean of precision and recall normalized between 0 and 1. Support: It is the number of actual occurrences of the class in the test data set.)

Table 2. ICBHI Respiratory infections detection (6-class) with Gaussian HMM Model



Column1	Precision	Recall	F1 Scores	Support
Broncheastasis	0.12	0.73	0.21	15
Broncholitis	0.04	0.89	0.08	27
COPD	0.99	0.28	0.44	1017
Healthy	1	0.02	0.04	49
Pneumonea	0.09	0.39	0.14	46
URTI	0	0	0	38
Accuraccy			0.28	1192
Macro Avg	0.37	0.39	0.15	1192
Weighted Avg	0.89	0.28	0.38	1192

Table 3. ICBHI Respiratory infections detection (6-class) with hybrid DNN-HMM Model



Column1	Precision	Recall	F1 Scores	Support
Broncheastasis	0.35	0.86	0.5	15
Broncholitis	0.15	0.65	0.12	27
COPD	0.88	0.57	0.76	1017
Healthy	0.2	0.26	0.4	49
Pneumonea	0.22	0.35	0.27	46
URTI	0.2	0.04	0.3	38
Accuraccy			0.56	1192
Macro Avg	0.28	0.44	0.29	1192
Weighted Avg	0.84	0.56	0.66	1192

According to the above comparison provided in Table 2 and Table 3, we can see that the disturbance in the amount of data. Table 1,2 summarizes that the accuracy of the Gaussian HMM model is 28% and for the DNN-HMM model accuracy is 56% So, looking into the results we can easily conclude that the overall accuracy of DNN-HMM is higher than that of the Gaussian HMM Model.

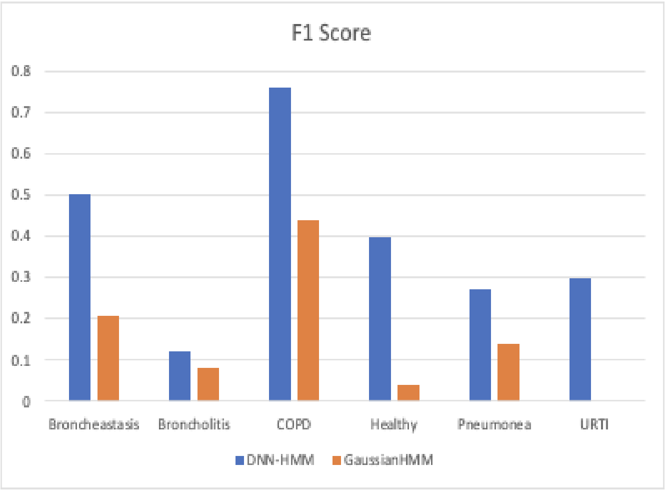


Fig. 5. Comparing F1 Score

Now, looking at the bar chart provided in Fig. 5 we can easily analyze that the F1 score which is the weighted harmonic mean of precision and recall and we can easily draw conclusions about the accuracy of individual classified diseases.

Firstly, the data distribution in the dataset amongst diseases is not balanced as we can see COPD has more data compared to other diseases. Gaussian HMM predicted poorly for every disease. Moreover, the model unable to detect UTRI and that's why there is no orange bar for UTRI.

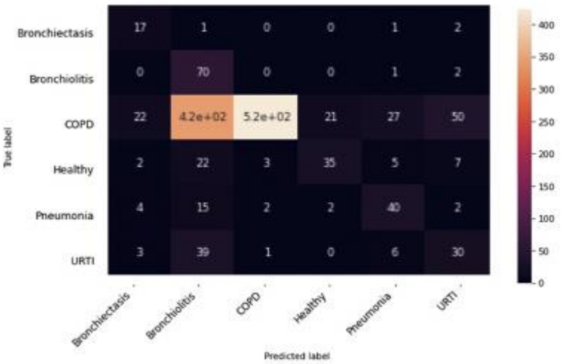


Fig. 6. Gaussian HMM Confusion Matrix

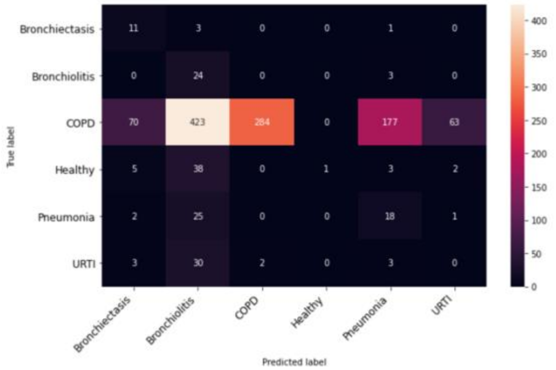


Fig.7. DNN-HMM Confusion Matrix

The confusion matrix provided in Fig. 6 and Fig. 7 for Gaussian HMM and DNN-HMM, respectively is the best way to compare the difference in true label and prediction label for all the 6 classes. The Y-axis is providing information on the true label and the X-axis is providing Information on the predicted label. We can conclude how many times are falsely predicted in each model and vice versa how many times the model made a correct prediction. The third axis represents the color according to the correctness of the prediction. If light color is max wrongly predicted and darkest color show least miss predicted. The confusion matrix is a highly efficient way to analyze the model results with a particular class.

VI. Conclusion and Future work

In this project, we developed a deep-learning framework along with HMM that originally integrates MFCC-based preprocessing of sound data and hybrid DNN-HMM feed-forward supervised models for the detection of six different respiratory diseases (crackles and wheezes) including the healthy lungs. Our empirical findings, drawn from an extensive evaluation conducted on the ICBHI challenge data and we have taken Gaussian HMM as a baseline model to compare the results. The hybrid DNN-HMM model performed quite well in comparison to the baseline model. The accuracy of DNN-HMM came 50%. If we would have classified it into normal and infectious the accuracy would have increased up to 90%. This is certainly due to the variable quality of the records, with sometimes a lot of background noise and interferences. The unbalanced data could also be a cause of bad predictions because it affects the model training. In the future, this model can be implemented to balanced and less noisy data for generating higher accuracy results.

VII. References

[1] "The global impact of respiratory disease (second edition)," Forum of International Respiratory Societies, 2017.

[2] A. A. Cruz, Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach. WHO, 2007.

[3] P. G. Burney, J. Patel, R. Newson, C. Minelli, and M. Naghavi, "Global and regional trends in copd mortality, 1990–2010," European Respiratory J., vol. 45, no. 5, pp. 1239–1247, 2015.

[4] "The global asthma report 2018," Global Asthma Network, 2018.

[5] T. Wardlaw, P. Salama, E. W. Johansson, and E. Mason, "Pneumonia: the leading killer of children," The Lancet, vol. 368, no. 9541, pp. 1048–1050, 2006.

[6] World malaria report 2015. World Health Organization, 2016.

[7] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," Cancer journal for clinicians, vol. 65, no. 2, pp. 87–108, 2015

[8] S. I. Khan and V. Ahmed, "Classification of pulmonary crackles and pleural friction rubs using MFCC statistical parameters," *2016* International Conference on Advances in Computing, Communications and Informatics *(ICACCI)*, 2016, pp. 2437-2440, doi: 10.1109/ICACCI.2016.7732422.

[9] Arts, Luca et al. "The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis." Scientific reports vol. 10,1 7347. 30 Apr. 2020, doi:10.1038/s41598-020-64405-6

[10] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," in *IEEE Access*, vol. 7, pp. 53040-53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[11] L. Li *et al*., "Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 312-317, doi: 10.1109/ACII.2013.58.

[12] https://arxiv.org/pdf/1907.05708.pdf.