Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp. 6926-6935
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.04.765098

Respiratory Lung Disease Classification Using Machine Learning Techniques

[1]Bharathy S , [2]Karthiga Priya N , [3]Meenaloshani K

[1]Assistant Professor, Department of Computer Scienceand Engineering Sri Sairam Engineering College

[2,3]Student, Department of Computer Science andEngineering Sri Sairam Engineering College

Abstract: At present, a large number of people lose their lives due to different respiratory diseases every day. Respiratory Sound Analysis has been a key tool to accurately detect these types of diseases. Earlier, manual detection of respiratory sounds was used but it is not feasible to detect various lung diseases due to various reasons like audio quality and perceptions of different doctors. Modern computer aided analysis helps to identify the diseases better with the sound and earlier treatment can be given to patients. These respiratory sound diseases include Asthma, Bronchitis, Pneumonia, COPD and URTI. The prediction with decision trees gives an accuracy rate of 88 percent, support vector machine gives an accuracy rate of the 82 percent and logistic regression gives an accuracy of 72 percent. A CNN model built and trained using the spectrogram images of audio files gave an accuracy rate of the 82 percent. Thus the proposed system detects the disease more earlier than manual detection with the help of respiratory sounds and tells us the exact lung disease in a better way.

Index terms: Convolution Neural Network.

INTRODUCTION

Respiratory sounds are one of the important signs of lungs health and respiratory disorders. The sound that is produced when a person breathes is directly associated to the movement of air, variations in the lung tissue and the position of the secretions inside lung. A wheezing sound is an example for a person with obstructive disease like asthma or chronic obstructive pulmonary disease (COPD).

COPD is marked by periodic exacerbation having symptoms of breathlessness and markable sputum production which worsen acutely, that results in hospitalisation. Cigarette smoking, air pollution, occupational exposure and aging are the factors that contribute COPD. The important treatment outcomes of COPD are symptoms, acute exacerbations and limitations of airflow

The relationship between detection of human pulmonary disorders and auscultating to respiratory lungs sounds is also discovered by Laennec. With the help of a stethoscope the

physician can detect signs of respiratory disorders and which enable them to diagnose the disease. However, there are many limitations in the application of stethoscope in research studies because of the variability in inter- observer and subjectivity in lung sounds interpretation. Diagnosis of the diseases from lung sounds needs professional training and experts.

In this context a technique that can automatically and accurately classify the sounds of lungs into many groups is very meaningful. It helps to detects potential threats at very early stage, even at home without a doctor. So, the most important purpose of the research is to develop an automated system to predict and diagnosis the respiratory diseases using lungs sounds. The most important objective of the research is to detect and categorize the lung noise digital signal with the help of signallearning processing methods. More specifically the study will be comparing the performance of convolutional neural network architecture and machine learning algorithms and to create a model based on the best performing algorithm.Ideally, this technique will improve detection of sounds and categorization of accuracy and robustness when encountered with different modes of sound and additional components while gaining the lung vibration wave.

II. LITERATURE REVIEW In [1] the model is able to achieve state of the art score on the ICBHI'17dataset. Deep learning models are shown to successfully learn domain specific knowledge when pre-trained with breathing data and produce significantly superior performance com-pared to generalized models. Local log quantization of trained weights is shown to be able to reduce the memory require- ment significantly. This type of patient-specific re-

training strategy can be very useful in developing reliable long-term automated patient monitoring systems particularly in wearable healthcare solutions.

In [2], an approach considered the breathing problems of patients as well as Asthma, Chronic Obstructive Pulmonary Disease (COPD), Tuberculosis, Pneumothorax and Lung cancer. Machine Learning and Deep Learning used to process data as well as create models for diagnosing patients. Combining the processing of patient information with data from chest X-rays, using CNN with the well-known pre-trained model, Caps Net network for data this form are the methods used for this project to identify the lung diseases. Initially studied and analyzed the data set, then apply Machine Learning and Deep Learning to predict that the patient has a lung disease or not. Project is a binary classification with input is patient's data (age,gender, chest X-ray images & view position) and output is found what the diseases is or not. The aim of the paper is to detect and diagnose the lung diseases as early as possible which will help the doctor to save thepatient's life.

In [3], machine learning algorithms; frequency coefficient features in a support vector machine (SVM) and spectrogram images in the convolutional neural network (CNN). Since using MFCC features with a SVM algorithm is a generally accepted classification method for

audio, we utilized its results to benchmark the CNN algorithm. We prepared four data sets for each CNNand SVM algorithm to classify respiratory audio:
(1) healthy versus pathological classification; (2) rale, rhonchus, and normal sound classification; (3) singular respiratory sound type classification; and
(4) audio type classification with all sound types. Accuracy results of the experiments were; (1) CNN 86%, SVM 86%, (2) CNN 76%, SVM 75%, (3) CNN 80%, SVM 80%, and (4) CNN 62%, SVM 62%, respectively. As a result, we found out that spectrogram image classification with CNN algorithm works as well as the SVM algorithm, and given the large amount of data, CNN and SVM machine learning algorithms can accurately classify and pre- diagnose respiratory audio.

III. SYSTEM ANALYSIS

1) EXISTING SYSTEM

Güler et al conducted a study by using power spectral density, normal, wheeze and crackles respiratory sounds are characterized with the help of power spectral density. The breathing sounds were obtained from subjects with various pulmonary diseases and also from healthy subjects. The breathing sounds are defined as having crackles, wheezes or normal respiratory sounds using GANN. A genetic algorithm is mainly implemented to choose neural network topology including upgrading of applicable element subsets and deciding the ideal number of hidden layers and handling elements. The component subsets, the

quantity of hidden layers, furthermore the quantity of handling components are the engineering factors of neural networks are to be resolved before the modelling process of neural network. Welch method is used for the spectral analysis of the respiratory sounds. A hybrid GANN approach was used to improve the prediction success rate of the network. This system is largely dependent on the amount of data and parameters used for training.

Chen et al proposed an efficient and automatic diagnostic method to analyze and categorize the lung sounds and heart sounds. They assembled the datasets of normal and abnormal sounds of lungs and professional doctors annotated the sounds. The lung sounds are categorized intonormal, wheezing rale and moist rales.

Bardou, Zhang and Ahmad acknowledge in their article about three method for classifying lung sounds. The first and second method were based on the extraction of a set of handcrafted features trained by three different classifiers (support vector machines, k-nearest neighbors, and Gaussian mixture models) while the third approach is based on the design of convolutional neural networks (CNN). From the audio files they extracted 12 MFCC coefficients after that six MFCCs statistics were calculated in the first approach. In order to increase the accuracy, they experimented normalization using zero mean and unity variance. While in the second approach visual representation of audio files are used to extract the local binary pattern (LBT) features. Using a whitening these features are normalized. The dataset were used contains 7 classes (normal, coarse crackle, fine crackle, monophonic wheeze,

polyphonic wheeze, squawk, and stridor). Tested dataset augmentation techniques on spectrograms are experimented to enhance the accuracyof CNN.

2) PROPOSED SYSTEM

Proposed system it mainly focus on the development of cost-effective and easy-to-use electronic device which can be applied in any device. 17,930 lung sounds from 1630 subjects can be recorded with the help of this device. They used two different types of machine learning algorithm for that, they are frequency coefficient features in a support vector machine (SVM) and spectrogram images in the CNN. For classification of the respiratory sounds they used four data sets for each CNN and SVM. The four data sets are healthy versus pathological, rhonchus, rale and normal sound, singular respiratory type and audio type with all sound type. Finally, they found that the image classification with CNN algorithm works as well as the SVM algorithm, and given the large amount of data, CNN and SVM machine learning algorithms can accurately classify and

pre- diagnose respiratory audio .The short-term spectral characteristics of lung sounds for the identification of associated diseases. They evaluated five different features to recognize three types of lung sounds: normal, wheeze and crackle.

With the help of lung sounds using wavelet coefficients and their classification by support vector machines and decision tree. It is found that decision tree has a higher than support vector machine performance based on which appropriate lung diseases are identified and results are obtained.

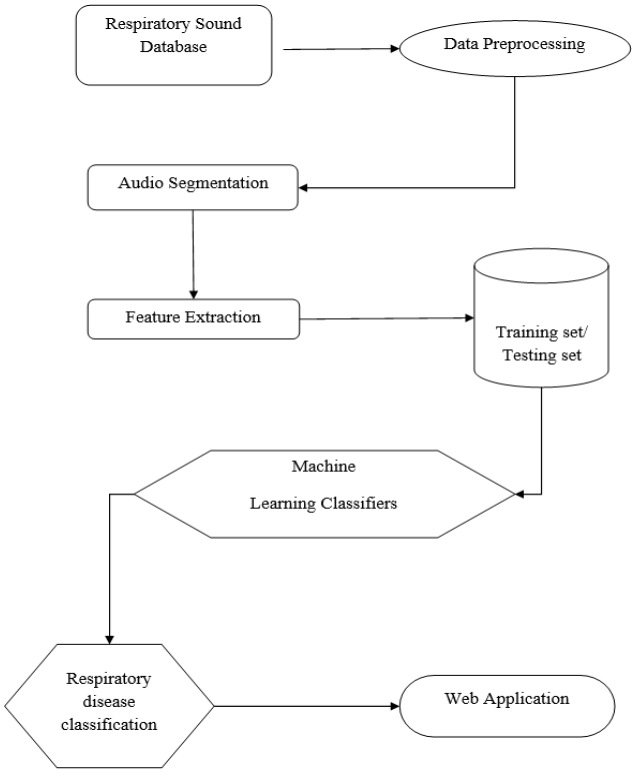

FIGURE 1: ARCHITECTURE OF RESPIRATORY LUNG DISEASE CLASSIFICATIONUSING MACHINELEARNING TECHNIQUES

IV. METHODOLOGY

The flowchart of various steps involved in the proposedsystem is

i DATA COLLECTION

The respiratory sounds were recorded utilizing a multi- channel or single channel obtaining technique, with each channel highlighting to an auscultating purpose of the member and each channel is saved in different files. The auscultation points are: Anterior right (Ar), Lateral Right (Lr), Posterior right (Pr), Trachea (Tc), Anterior left (Al), Lateral left (Ll) and Posterior left (Pl). The heterogenous equipment 's used to acquire respiratory sounds by Lab3R are: "Welch Allyn Meditron Master Elite Plus Stethoscope Model 5079-400" digital stethoscope Seven "3M Littmann Classic II SE" stethoscopes with a microphone in the main tube Seven air coupled electret microphones (C 417 PP,

AKG Acoustics) locatedin capsules made of Teflon.ii RESPIRATORY SOUND DATABASE
The programmed examination of respiratory sounds has been a field of incredible research enthusiasm during the most recent decades. Computerized characterization of respiratory sounds can possibly distinguish irregularities in the beginning times of a respiratory brokenness and along these lines improve the adequacy of dynamic. Be that as it may, the presence of a publicly accessible enormous database, in which new calculations can be executed, assessed, and thought about, is as yet missing and is indispensable for additional improvements in the field. The making of this database and the related logical test comprise an underlying yet unequivocal advance towards utilizing computational lung auscultation, and furthermore towards featuring the intricacy of the Respiratory Sound characterization issue.

The database for Respiratory Sound was initially

breathing portion of the audio file, we wish to slice the wav file into sub slices. Again, the start and end times given in the .txt files indicate this. Items we do need to remember when we slice the audio files. We need to make sure they have the same duration (this is in preparation for subsequent feeding them into the training model) If they are not the same length, then we have to pad silent (or zero) soundsto the audio. This has separate Stereo and Mono equations.
Stereo: (sampling rate time)
*2 Mono: (Sampling rate *time)

iv HANDLING MISSING DATA

A preprocessing method for handling the missing data is done. The data files are read which gives the demographic information, patient details and combine them to create a new preprocessed data. At the first we check the total number of the missing values using the is null (). sum ().Once we get the total missing values, we eliminate any rows that has 3 missing values or more. After the missing values are removed, from the data created to support the research challenge that was available for us the Body Mass index of the patient is not organized at

Int. Conf. On

MachineH

research groups in two distinct nations, more than quite a long while. The vast majority of the database comprises of sound examples recorded by the School of Health Sciences, University of Aveiro (ESSUA) inquire about group at the Respiratory Research and Rehabilitation Laboratory (Lab3R), ESSUA and at Hospital Infante D. Pedro, Aveiro, Portugal. The subsequent research group, from the Aristotle University of Thessaloniki (AUTH) and the University of Coimbra (UC), gained respiratory sounds at the Papanikolaou General Hospital, Thessaloniki and at the General Hospital of Imathia (Health Unit of Naousa), Greece.

The database comprises of an aggregate of 5.5 long stretches of recordings containing 6898 breathing cycles, of which 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes, in 920 commented on sound examples from 126 subjects.

The cycles were commented on by respiratory specialists as including crackles, wheezes, a blend of them, or no extrinsic respiratory sounds. The accounts were gathered utilizing heterogeneous hardware and their term ran from 10s to 90s. The chest areas from which the accounts were gained is additionally given. Noise levels in some breath cycles is high, which reproduce genuine conditions.

iii PREPROCESSING

First load the patient diagnosis file and check all the specific diagnosis that we have in our records. This is key in the way we sort our performance later. We will need toread all of the specific files in our dataset next. This is achieved using the function os.listdir with the condition that only.txt files are checked. Now that we'vegot our list of files, we have to read each one to get the details about crackles and wheezes including when it's captured in the audio file (start and end time in seconds). To get the pure available for some patients we use the mean of the BMI toadd the missing data's to the rows.

V. RESULT

Implementing a system for the respiratory sound classification and prediction using various machine learning algorithms this helps us to identify which algorithms give better performance of the prediction .We use the python as a background for our development of the system as it gives more functionality for data analysis. By importing the necessary libraries. Load the train and test data using the flow from directory Image Data Generator. Target size

set to 224x224. Using Image Net weights to get the model pre trained. The performance prediction is set to 8 since we have 8 classes

.Use Adam as our optimizer and set the learning rate to 0.000001 as we try to identify signals again. Even it's not a regular sound wave either since we use breathing sounds. Therefore, the model must be very sensitive to the small details .Even, if the validation accuracy increases, we want to save the pattern. But if we see the lack of improvement after 20 epochs, we do not want to waste computational resource. We then describe the checkpoint for our model and the early stoppage. In view of the imbalance of our data, we have to calculate the class weight for each category class. We have calculated steps per epoch and validation steps instead of simply setting a random value. The explanation is that we need to look for the best way to get these values. This is calculated by dividing the samples batch size.

i Decision Tree

The accuracy and the classification report is as follows. the precision, recall and f1-score and the global accuracy of the model. The weighted average of the f1 has to be considered in classification problems rather than the accuracy. Here we get 88 percent as the weighted average and that show a good performance for the algorithm.

```
Accuracy: 0.90
Confusion Matrix:
 [[ 3  0  0  0  0  0]
 [ 1  4  0  0  0  0]
 [ 0  0 15  0  0  0]
 [ 0  0  0  1  0  0]
 [ 0  0  1  0  3  0]
 [ 1  0  0  0  0  0]]
              precision    recall  f1-score   support

           0       0.60      1.00      0.75         3
           1       1.00      0.80      0.89         5
           2       0.94      1.00      0.97        15
           3       1.00      1.00      1.00         1
           4       1.00      0.75      0.86         4
           5       0.00      0.00      0.00         1

    accuracy                           0.90        29
   macro avg       0.76      0.76      0.74        29
weighted avg       0.89      0.90      0.88        29
```

FIGURE NO 2: ACCURACY OF DECISION TREE

ii Support Vector Machine Algorithm

The SVM classifier gives an overall accuracy rate of 83 percent in the testing set with and weighted f1 average of 76 percent.

```
SUPPORT VECTOR MACHINE ALGORITHM

Accuracy: 0.83
Confusion Matrix:
 [[ 2  1  0  0  0  0]
 [ 0  5  0  0  0  0]
 [ 0  0 15  0  0  0]
 [ 0  0  0  1  0  0]
 [ 0  0  4  0  0  0]
 [ 0  0  0  0  0  1]]
              precision    recall  f1-score   support

           0       1.00      0.67      0.80         3
           1       0.83      1.00      0.91         5
           2       0.79      1.00      0.88        15
           3       1.00      1.00      1.00         1
           4       0.00      0.00      0.00         4
           5       1.00      1.00      1.00         1

    accuracy                           0.83        29
   macro avg       0.77      0.78      0.77        29
weighted avg       0.72      0.83      0.76        29
```

FIGURE NO 3: ACCURACY OF SVM ALGORITHM

iii Logistic Regression

In logistic regression on the end gives an overall accuracy of the 72 percent and anf1 weighted average score of 63 percent, this seems to be worst performingalgorithm from all the above.

```
Accuracy: 0.72
Confusion Matrix:
 [[ 1  2  0  0  0  0]
 [ 0  5  0  0  0  0]
 [ 0  0 15  0  0  0]
 [ 0  0  1  0  0  0]
 [ 0  0  4  0  0  0]
 [ 1  0  0  0  0  0]]
              precision    recall  f1-score   support

           0       0.50      0.33      0.40         3
           1       0.71      1.00      0.83         5
           2       0.75      1.00      0.86        15
           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00         4
           5       0.00      0.00      0.00         1

    accuracy                           0.72        29
   macro avg       0.33      0.39      0.35        29
weighted avg       0.56      0.72      0.63        29
```

FIGURE NO 4:ACCURACY OF LOGISTIC REGRESSION

Algorithms Comparison

Among the algorithms of machine learning and from the CNN model compared the best performing algorithm is Decision tree with 90 percent accuracy and we use this for the prediction of respiratory diseases. And also, from the comparing the algorithms we can identify that logistic regression isperforming the leas

FIGURE NO 5: DETECTION OF COPD DISEASE

VI . CONCLUSION

Using various Machine Learning algorithm the respiratory diseases is predicted from the respiratory sounds database using decision trees. This comprised of the comparison with various machine learning algorithms. The system mainly consists of three areas, preprocessing of the data, prediction of diseases, and developing the interface for users to use .In the preprocessing, we are handling the missing data, normalizing the values and eliminating any unwanted data from our dataset and creating a new preprocessed data. The prediction with decision trees gives an accuracy rate of 90 percent, support vector machine gives an accuracy rate of the 83 percent and logistic regression gives an accuracy of 72 percent. A CNN model built and trained using the spectrogram images of audio files gave an accuracy rate of the 82 percent. CNN models can be used when there is a large amount of data is available to train, when learning with less data the CNN model is not so appropriate as overfitting occurs in the training of model. The more the accuracy and f1 average weight of the algorithms the model's predictions become accurate.

REFERENCE

[1] Acharya, Jyotibdha, and Arindam Basu. "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning." IEEE transactions on biomedical circuits and systems 14.3 (2020): 535-544.

[2] Demir, Fatih, Abdulkadir Sengur, and Varun Bajaj. "Convolutional neural networks based efficient approach for classification of lung diseases." *Health information science and systems* 8.1 (2020): 1-8.

[3] Glos, Martin, et al. "Tracheal sound analysis for detection of sleep disordered breathing." Somnologie 23.2 (2019): 80-85.

[4] Henry, Brian, and Thomas J. Royston. "Localization of adventitious respiratory sounds." The Journal of the Acoustical Society ofAmerica 143.3 (2018): 1297- 1307.

[5] Aykanat, Murat, et al. "Classification of lung sounds using convolutional neural networks." EURASIP Journal on Image and Video Processing 2017.1 (2017): 1-9.

[6] Rocha, B. M., et al. "A respiratory sound database for the development of automated classification." International Conference on Biomedical and Health Informatics. Springer(2017).

[7] Cordel II, Macario O., and Joel P. Ilao. "A Computer Assisted Diagnosis System for the Identification/Auscultation of Pulmonary Pathologies." Manila Journal of Science 9 (2016): 8-26.

[8] Chen, Qiyu, et al. "Automatic heart and lung sounds classification using convolutional neural networks." (2016)Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 201