Sound Classification for Respiratory Diseases Using Machine Learning Technique

Dr.B.Bharathi Kannan., S.Pushparani
M.Tech.,PHD, Head of the Department Assistant Professor, Dept. Of Computer Science and Engineering, M.I.E.T Engineering College, Trichy
Me Student, Master of Engineering in Computer Science and Engineering, M.I.E.T Engineering College, Trichy

Abstract: **Respiratory sounds are one of the important signs of lung health and respiratory disorders. These respiratory sounds can be acquired using digital stethoscopes and other recording devices. This advanced information opens up the chance of utilizing artificial intelligence to naturally analyze respiratory scatters like asthma, pneumonia and bronchiolitis, to give some examples. A very high number of people lose their lives 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 give much better results in identifying the diseases from the sound that includes identification of Wheezes and Crackles in the audio and thus better treatment can be given to patients. These respiratory sound diseases include Asthma, Bronchitis, Pneumonia, COPD (Chronic Obstructive Pulmonary Disease), LRTI (Lower Respiratory Tract Infection), and URTI (Upper Respiratory Tract Infection).**

Keyword: Lung diseases, Digital stethoscopes, Sound analysis, Artificial Intelligence, Wheezes, Crackles

1. INTRODUCTION

The classification and identification of breathing diseases is a tedious task. The sound that is produced when a person breathes is directly associated with the movement of air, variations in the lung tissue and the position of the secretions inside the lung. A wheezing sound is an example for a person with obstructive disease like asthma or Chronic Obstructive Pulmonary Disease (COPD). One of the major causes of mortality and morbidity worldwide is respiratory diseases. It developed the third prominent cause of death in 2020. Asthma is also related to COPD, but the definition is different. This disease also results in social and economic burden that is both substantial and increasing. The important treatment outcomes of COPD are symptoms, acute exacerbations and limitations of airflow. Interestingly, the Sounds from the lungs conveys significant information associated with respiratory diseases and it helps to assess the patients with pulmonary or respiratory disorders. Sounds released from a person's breath are directly related to changes in lung tissue, position of secretion within the lungs and air movement. For instance, wheezing sound is a common indication that the patients have diseases like obstructive airway disease (asthma and chronic obstructive pulmonary disease).The most important objective of the research is to detect and categorize the lung noise digital signal with the help of signal learning processing methods. An electronic stethoscope overcomes the low sound levels by electronically amplifying the body sounds. Electronic stethoscopes convert the acoustic sound waves obtained through the chest piece into electrical signals which may then be amplified for optimal listening. Recording equipment includes AKG C417L Microphone, 3M Littmann Classic II SE Stethoscope, 3M Littmann 3200 Electronic Stethoscope, WelchAllyn Meditron Master Elite Electronic Stethoscope. In this paper we are focusing on the respiratory sound classification and prediction using machine learning algorithm (Decision Tree Classifier).We use the python as a background for our development of the system as it gives more functionality for data analysis.

2. LITERATURE SURVEY

In this paper [1] The stethoscope and its semantic of auscultatory findings were invented 200 years ago by Dr. Laennec and over the years only a few changes were made to both the stethoscope the way during which it is used. However, the process to distinguish between normal and abnormal sounds or noises (vesicular sounds, wheezes, crackles, etc.) remains

essential in clinical practice for proper diagnosis and management. It aims to review recent technological advances, evaluate promising innovations and perspectives within the field of auscultation, with a special intelligent communicating stethoscope systems in clinical practice, and within the context of teaching and telemedicine

In this paper [2] Identification of normal and abnormal respiratory sounds such as crackles, wheezes is very essential for accurate diagnosis of diseases. These sounds include a lots of information about the pathologies and physiologies of lung structure and any obstruction in airways can be identified from the sounds.

In this paper [3] Various studies were done, and research was made to test human ears' capacity to identify crackles. The research consisted of crackles simulated to superimpose as real respiratory/breathing sound. The most important detection errors were identified from these research like intensity of crackles, different types of crackles make different wavelengths and so on. From these studies we could conclude that traditional auscultation should not be considered as individual reference for validating respiratory sounds.

In this paper [4] Supervised machine learning is the construction of algorithms that are able to produce general patterns and hypotheses by using externally supplied instances to predict the fate of future instances. Supervised machine learning classification algorithms aim at categorizing data from prior information. Classification is carried out very frequently in data science problems. Various successful techniques have been proposed to solve such problems viz. Rule-based techniques, Logic-based techniques, Instance-based techniques, stochastic techniques. This paper discusses the efficacy of supervised machine learning algorithms in terms of the accuracy, speed of learning, complexity and risk of overfitting measures. The main objective of this paper is to provide a general comparison with state of art machine learning algorithms.

In this paper [5] "COPD has been perceived as being a disease of older men. However, >7 million women are estimated to live with COPD in the USA alone. Despite a growing body of literature suggesting an increasing burden of COPD in women, the evidence is limited. To assess and synthesize the available evidence among population-based epidemiologic studies and calculate the global prevalence of COPD in men and women. A systematic review and meta-analysis reporting gender-specific prevalence of COPD was undertaken. Gender-specific prevalence estimates were abstracted from relevant studies. Associated patient characteristics as well as custom variables pertaining to the diagnostic method and other important epidemiologic covariates were also collected. A Bayesian random-effects meta-analysis was performed investigating gender-specific prevalence of COPD stratified by age, geography, calendar time, study setting, diagnostic method, and disease severity. Among 194 eligible studies, summary prevalence was 9.23% (95% credible interval [CrI]: 8.16%–10.36%) in men and 6.16% (95% CrI: 5.41%–6.95%) in women.We conducted the largest ever systematic review and meta-analysis of global prevalence of COPD and the first large gender-specific review. These results will increase awareness of COPD as a critical woman's health issue.

3. METHODOLOGY

3.1 EXISTING SYSTEM

The existing system mainly focus on the finding the respiratory sounds that has the potential to detect abnormalities in the early stages of a respiratory dysfunction and thus enhance the effectiveness of decision making. However, the existence of a publically available large database, in which new algorithms can be implemented, evaluated, and compared, is still lacking and is vital for further developments in the field. The recordings were collected using heterogeneous equipment and their duration ranged from 10 to 90 s. The chest locations from which the recordings were acquired was also provided. Some disadvantages are method is costly, can be used only to predict only 2 respiratory lung diseases, noise level is high, frequently the x-rays need to be taken for prediction.

3.2 PROPOSED SYSTEM

The proposed method uses the identification of the respiratory disease with the help of the datasets that consists of the people with breathing problems. It identifies and tells us the exact respiratory problem that occurs for the individuals. The prediction is exact and the algorithm used are efficient in finding the respiratory disease. This helps in knowing the problem prior to the last stage.

3.3 ALGORITHM - XG BOOST

Boost stands for Extreme Gradient Boosting. It uses more accurate approximations to find the best tree

model. Boosting: N new training data sets are formed by random sampling with replacement from the original dataset, during which some observations may be repeated in each new training data set.For each node, there is a factor γ with which hm(x) is multiplied. This accounts for the difference in impact of each branch of the split. Gradient boosting helps in predicting the optimal gradient for the additive model, unlike classical gradient descent techniques which reduce error in the output at each iteration.

GBoost Features

Regularized Learning:
The regularization term helps to smooth the final learned weights to avoid over-fitting. The regularized objective will tend to select a model employing simple and predictive functions.

Gradient Tree Boosting:
The tree ensemble model cannot be optimized using traditional optimization methods in Euclidean space. Instead, the model is trained in an additive manner.

Shrinkage and Column Subsampling:
Besides the regularized objective, two additional techniques are used to further prevent overfitting. The first technique is shrinkage introduced by Friedman. Shrinkage scales newly added weights by a factor η after each step of tree boosting. Similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each tree and leaves space for future trees to improve the model.

XGBoost works step by step
Image result for xgboost algorithm pseudocode
Step 1: Create an Respiratory sounds Dataset.
Step 2: Create a PHP Traning dataset.
Step 3: Mysql Transform Data.
Step 4: Train a Model.
Step 5: Deploy the Model.
Step 6: Evaluate the Model.
Step 7: Clean Up.

Random forest
Random forest is a accurate method. It is used in two classifications. That two method is supervised and learning algorithm. Classification a problem has solved by these methods .This method lot of trees is formed this forest method. From the data samples a decision tree will be introduced in this forest method. Use voting to choose simple result of solution.

Working of Random Forest Algorithm
Step 1 – In given data set, we can select the random samples
Step 2 – Decision tree will be farming for each sample. It will show prediction result
Step 3 – Voting will be started
Step 4 – Final result is most voted prediction

Linear regression
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Disease Prediction Algorithm:
Step 1: Reading the pre-processed data csv file.
Step 2: Define variables X and Y from dataset.
Step 3: Encoding categorical data e.g. gender as a dummy variable.
Step 4: Encoding categorical data e.g. disease outcome as a dummy variable.
Step 5: Splitting the dataset into the Training set and Test set.
Step 6: Fitting Classifier to the Training Set (Decision Tree Classifier)
Step 7: if request.method == 'POST':
A = request.form['Age']
N = request.form['Gender']
P = request.form['BMI']
Step 8: Creating a function for Annotation data for identifying recording_info and recording_annotations.
Step 9: Summed number of crackles / wheezes are normalized by the duration of the recording
duration = annotation.iloc[-1, 1] - annotation.iloc[0, 0]
info['Crackles'] = crackles/duration # crackles per second
info['Wheezes'] = wheezes/duration # wheezes per second
Step 10: prediction = classifier.predict(pred)
if prediction==5:
print('Bronchiolitis')
if prediction==4:
print('Pneumonia')
if prediction==3:
print('Bronchiectasis')
if prediction==2:
print('COPD')
if prediction==1:

print('Healthy')
if prediction==0:
print('URTI').

3.4 ARCHITECTURE DIAGRAM

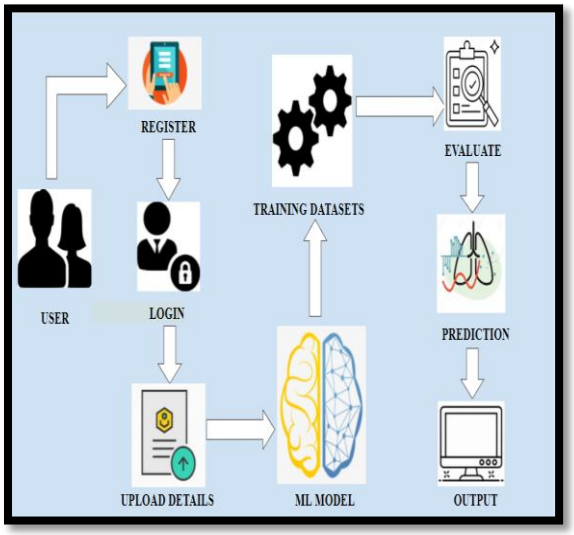


Fig : System architecture

4. EXPERIMENTAL RESULTS

The base idea is to create a web application for the prediction of respiratory diseases. Initially the patient's data and their corresponding respiratory sounds are collected using various digital stethoscopes. The respiratory sounds are in .wav format that are observed at several patient's chest locations such as Trachea, Anterior, Posterior, Lateral. 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. Once we get the total missing values, we eliminate any rows that have 3 missing values or more. After the missing values are removed, from the data available for us the Body Mass Index of the patient is not available for some patients we use the mean of the BMI to add the missing data's to the rows. Once the missing data is handled, we now read the information audio from the txt file corresponding to each audio file and it creates two data frames one with

4.1 IMPLEMENTATION

Admin Module

This module is used to login for administrator, it have whole rights to monitor and manage the entire project through this module. Admin can train the different datasets which is taken from the KAGGLE website. The training phase datasets consists of age, gender, breath rate etc. Admin can view all the patient details and their relevant data. The result is the screen that displays the chance of a person having respiratory disease. Then produce counselling and certain treatment for the person who is affected by the disease.

User module

With the datasets already required from the UCI repository, then the user details will be collected and store in the database. Data like the patients Name, Age, Gender, Patient No, BMI, Disease, Crackles, Wheezes . Already registered user can directly start accessing the system with the help of the user id and the password provided. The user can view the result consists diagnosis information and precaution details

Data acquisition

The process of data acquisition involves searching for the datasets that can be used to train the Machine Learning models. Data acquisition meaning is to collect data from relevant sources before it can be stored, cleaned, pre-processed, and used for further mechanisms. It is the process of retrieving relevant business information, transforming the data into the required business form, and loading it into the designated system.In this module the data is collected in the form of audio, these audio files are stored and they are converted in the form of a text file, this is how the data is collected. These data are collected from different patients' lungs and it is recorded. Data virtualization of audio and text files.

Pre-processing

Data pre-processing involves the transformation of the raw dataset into an understandable format. Pre-processing data is a fundamental stage in data mining to improve data efficiency. The data pre-processing methods directly affect the outcomes of any analytic algorithm. These data are pre-processed and we have to clean the unwanted values all the null values.

Features selection

Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. A related term, feature engineering (or feature extraction), refers to the process of extracting useful information or features from existing data. The features selection is done using Filter Method. The features are filtered based on general characteristics (some metric such as

correlation) of the dataset such correlation with the dependent variable. Filter method is performed without any predictive model. It avoids over fitting. It apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often uni-variate and consider the feature independently, or with regard to the dependent variable. It can be used to construct the multiple respiratory diseases. In this module, select the multiple features from uploaded datasets. And train the datasets with various disease labels such as COPD, URTI, LRTI, asthma, wheezes.

Classification

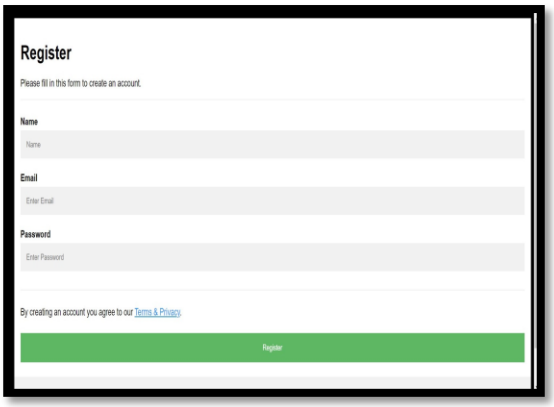
XGBoost (eXtreme Gradient Boosting) is a powerful machine learning algorithm that is mainly used for classification tasks. Collect the pre-process breath sounds . Divide the data into training, validation and testing sets. Define the architecture of the XGBoost model using the training set. validate the XGBoost model using the validation set. Test the XGBoost model using the testing set and determines the accuracy in classifying the breath sounds. The basic idea behind boosting is to combine a set of weak learners into a strong learner. Each decision tree is trained on a subset of the data and is designed to correct the errors made by the previous tree in the ensemble.

Disease diagnosis

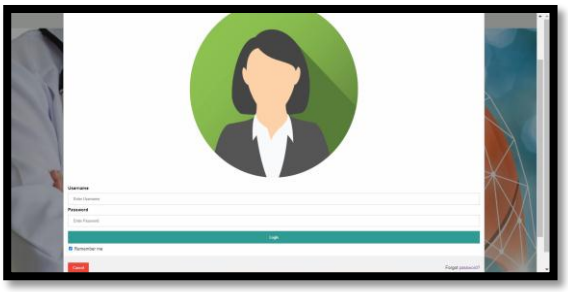
Medical decision support system is a decision-support program which is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patients' data. In this module, provide the diagnosis information based on predicted respiratory diseases. Proposed system provides improved accuracy in respiratory disease prediction. Risk factors are conditions or habits that make a person more likely to develop a disease.

4.2 RESULTS

Home Page

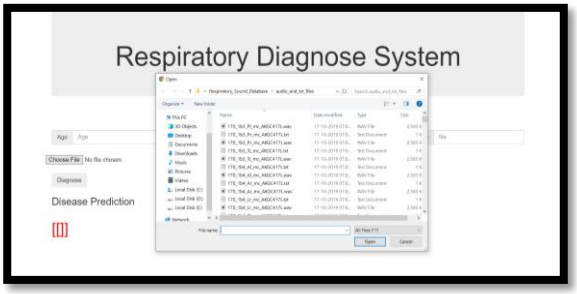


Register page



User Login

Disease Prediction using Wheezes and Crackles



Uploading Audio file for Prediction

Disease Prediction using Audio file

5. CONCLUSION

In our project, we have predicted respiratory diseases from the respiratory sounds database using decision trees. Our work comprised of the comparison with machine learning algorithm. Our project mainly consists of three areas, preprocessing of the data, prediction of diseases, and developing the interface for users to use. In the reprocessing, 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. 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 as appropriate as over fitting occurs in the training of model. The more the accuracy and f1 average weight of the algorithms the model's predictions become accurate.

FUTURE WORK

The future work of the application consists of mostly collecting more data and trying to implement with the CNN model, also instead of manual annotation of the audio files teach the model to automatically annotate the recordings. This web application can add storage functionalities where the users can access their previous breath sound checks and also the automatic annotation process of sounds which helps the users to easily identify the disease. A desktop or mobile application can also be built to make the process easier for the users.

REFERENCES

[1] Welsby, P.D., Parry, G. and Smith, D. 'The Stethoscope: Some Preliminary Investigations'. Postgraduate Medical Journal (2003)

[2] Swarup, S. and Makaryus, A.N. 'Digital Stethoscope: Technology Update'. Medical Devices (2018)

[3] Sandra Reichert, Gass Raymond, Christian Brandt, Emmanuel Andres 'Analysis of Respiratory Sounds: State of the Art'. Clinical Medicine. Circulatory, Respiratory and Pulmonary Medicine, (2008)

[4] H Kiyokawa, M Greenberg, K Shirota, H Pasterkamp 'Auditory detection of simulated crackles in breath sounds' Chest. 2001 Jun;119(6):1886-92.doi: 10.1378/chest.119.6.1886.(2001)

[5] Zimmerman, B. and Williams, D. 'Lung Sounds'. In StatPearls. Treasure Island (FL): StatPearls Publishing. Available at: http://www.ncbi.nlm.nih.gov/books/NBK537253/ . (2020a)

[6] Zimmerman, B. and Williams, D 'Lung Sounds'. In StatPearls. Treasure Island (FL): StatPearls Publishing. Available at: http://www.ncbi.nlm.nih.gov/books/NBK537253/ (2020b)

[7] Emmanuel Andres, Amir Hajjam. 'Advances and Perspectives in the Field of Auscultation, with a Special Focus on the Contribution of New Intelligent Communicating Stethoscope Systems in Clinical Practice, in Teaching and Telemedicine'.EHealth and Remote Monitoring. InTech. (2012)

[8] B. M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, R. P. Paiva, I. Chouvarda, P. Carvalho, N. MaglaverasA. 'Respiratory Sound Database for the Development of Automated Classification' Precision Medicine Powered by PHealth and Connected Health. IFMBE Proceedings. Singapore: Springer Singapore(2018)

[9] Renars Xaviero Adhi Pramono, Stuart bowyer, Esther Rodriguez-Villegas 'Automatic Adventitious Respiratory Sound Analysis: A Systematic Review' (2017)

[10] Singh, A., Thakur, N. and Sharma, A. 'A Review of Supervised Machine Learning Algorithms'3rd International Conference on Computing for Sustainable Global Development (INDIACom) (2016)

[11] Georgios Ntritsos, Jacob Franek, Lazaros Belbasis, Maria A Christou, Georgios Markozannes, Pablo Altman, Robert Fogel, Tobias Sayre,Evangelia E Ntzani, Evangelos Evangelou. 'Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis' International Journal of Chronic Obstructive Pulmonary Disease (2018).