KLE Society's
KLE Technological University

A Mini Project Report
On

Automated Heterogeneous Machine Learning Techniques for Respiratory Disease Analysis

Bachelor of Engineering
In
Computer Science and Engineering

Submitted By

Shreyas N B	**01FE19BCS005**
Vighnesh Pai	**01FE19BCS013**
Vishwanath Kammar	**01FE19BCS015**
Ramkrishna Mutalikdesai	**01FE19BCS024**
Sharath Shanbhag	**01FE19BCS029**

Under the guidance of
UMADEVI F M

.

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBLI–580 031 (India).

Academic year 2021-22

KLE Society's
KLE Technological University

2021 - 2022

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Mini Project entitled Automated Heterogeneous Machine Learning Respiratory Disease Analysis is a Bonafide work carried out by the student team Mr. Vighnesh N Pai – SRN 01FE19BCS013, Mr. Vishwanath Kammar – SRN 01FE19BCS015, Mr. Shreyas N B – SRN 01FE19BCS005, Mr. Ramkrishna Mutalikdesai – SRN 01FE19BCS024, Mr. Sharath Shanbhag – SRN 01FE19BCS029, in partial fulfillment of completion of Fifth semester B. E. in Computer Science and Engineering during the year 2021 – 2021. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the above said programme.

Guide **Head, SoCSE**
Ms. Uma Devi FM **Dr. Meena S. M**

External Viva:

Name of the Examiners **Signature with date**

1.
2.

\

ACKNOWLEDGEMENTS

In the accomplishment of this project successfully, many people have best owned upon us and pledged support. We express our sincere gratitude towards all the people who have been concerned with this project.

We take this opportunity to thank Dr.Ashok Shettar, Vicechancellor,KLE Technological University and to Dr. Prakash Tewari, Principal, B V Bhoomaraddi College of Engineering and Technology, Hubli.

We take this opportunity to thank Dr.Meena S M, Head, School of Computer Science and Engineering for having provided us with an academic environment which nurtured our practical skills contributing to the success of our project.

We sincerely thank our guide Ms. Uma Devi FM, Associate Professor, School of Computer Science and Engineering for her guidance, inspiration and wholehearted cooperation during the course of completion. We owe our gratitude to the School of Computer Science Engineering, KLE technological University for providing us with the resources necessary for the completion of this project. Our gratitude will not be complete without thanking the Almighty, our beloved parents, our seniors and our friends who have been a constant source of blessings and aspirations.

(Team members)

Vighnesh N Pai
Vishwanath Kammar
Shreyas N B
Ramkrishna Mutalikdesai
Sharath Shanbhag

INTRODUCTION

1. Introduction

1.1 Preamble

Respiratory disease is the most vulnerable disease found (in recent time) in the world and the reason behind the death of human beings or reduction in mortality rate. Detection and diagnosis of this deadly disease is the more challenging part. Covid-19, can be diagnosed in the early stages and can be treated early. Now a days covid-19 disease is the most vulnerable disease found in 2020

COVID-19 is a large-scale contagious respiratory disease that has spread across the world in 2020. Therefore, a low-cost, fast, and easily available solution is needed to provide a COVID-19 diagnosis to curb the outbreak. According to recent studies, one of the main symptoms of COVID-19 is coughing. The goal of this research effort is to develop a method for the automatic diagnosis of COVID-19 by detecting cough during recorded conversations. The method is composed of the five main modules: sound extraction, sound feature extraction, cough detection, cough classification, and COVID-19 diagnosis. The method extracts relevant features from the audio signal and then uses machine learning and deep learning models, like SVM, KNN, and RNN, to classify them, which can successfully diagnose from audio recordings. When dealing with completely unfamiliar cough samples. As more data sets are collected, the model can be further developed and improved to create a machine learning solution based on cough analysis for COVID-19, Asthma, Chronic obstructive pulmonary disease (COPD), Pneumonia detection which may be promoted as a non-clinical self-inspection solution. Chronic respiratory diseases, such as asthma and chronic obstructive pulmonary disease, usually kill more than four million people every year and affect hundreds of millions more around the world.

Women and children are particularly vulnerable, especially those in low and middle-income countries, where they are exposed on a daily basis to indoor air pollution from solid fuels for cooking and heating. In high-income countries, tobacco is the most important risk factor for chronic respiratory diseases, and in some of these countries, tobacco use among women and young people is still increasing.

1.2 Motivation

- Respiratory disease is becoming a major global health problem. With the machine learning of research around the world, more and more studies have shown that respiratory disease spreads from person to person through droplets or direct contact. According to recent research, the load of this virus is at its peak in the early stages of the illness, making respiratory disease infectious even before the symptom onset, companies around the world have developed a different method to detect the disease, but availability is limited.

- At present, the detection of respiratory disease is mainly conducted by medical staff and patients through a face-to-face collection of biological samples. This method has many disadvantages, such as spread risk, high detection cost, and long detection cycle.

- The cost of testing is more for common people. In some underdeveloped areas, some of the respiratory diseases are difficult to diagnose because of the virus due to lack of medical facilities and medical professionals. Therefore, a low-cost, fast, and easily available solution is needed.

1.3 Objectives

1.**Literature** survey study.

2.**Build** pipeline for using machine learning methods.

3.Selection of **appropriate features**.

4.**Evaluating** and analyzing the model.

5.**Compare** with existing works for further improvement.

1.4 Literature Survey

Title, Author, Year of Publication	**Method/Model**	**Performance**	**Remark**
1)An adaptive speech signal processing. -Kawther A. Al-Dhlan 2021	Generative Adversarial Network (GAN) classifier technique.	GAN: Precession- 96.54% Accuracy- 98.56% F-measure- 0.96	Generative Adversarial Network for detection of Covid-19.
2)Recognition of Pulmonary diseases using CNN and BDLSTM units -M. Fraiwan 2021	CNN+BDLSTM using Deep learning model	Average accuracy - 99.62% Cohens Kappa value – 98.26%	CNN and BDLSTM used for lung sound classification.

3)Deep Learning Based Respiratory Sound Analysis For detection of COPD Arpan Srivastava 2021	CNN [MFCC] method. (Mel Frequency Cepstral Coefficient)	Sensitivity- 0.93 Specificity- 0.93 ICBHI Score- 0.93	CNN based deep learning assistive model for detecting COPD
4)Multimodal point of care diagnostics for Covid-19 based on acoustic and symptoms Srikant Raj Chetupalli 2021	Multimodal diagnostic tool.	Specificity – 0.95 Accuracy – 92.7%	Simple linear models for covid diagnostics using multi-modal data
5)PANACEA cough and sound-based diagnosis of COVID-19 -Madhu R Kamble 2021	Multi-scale Domain-adversarial Multiple-instance CNN	AUC (Area Under Curve) - 76.31%	TECC and LightGBM features used in diagnosing of covid data.

1.5 Problem Definition

Automated heterogeneous machine learning methods to analyze respiratory diseases

- The project aims to analyze respiratory diseases using cough breath and speech samples. This approach is a non-invasive approach as the person's cough, breath and sound samples are processed. We will be taking different cough sound samples of infected people from Coswara, Coughvid dataset.

PROPOSED SYSTEM

2. Proposed System

2.1 Description of Proposed System

- The method comprises of five main modules: preprocessing, feature_extraction
- This method extracts relevant features from the audio signal and then deep learning models later to classify them.
- Which can successfully diagnose respiratory disease from audio recordings. Our method is relative when dealing with completely unfamiliar cough samples. When the training set and the test set are from two different databases.
- As more data sets are collected, the model can be further developed and improved to create a machine learning solution based on cough analysis for respiratory disease detection, which may be promoted as a non-clinical self-inspection solution.

2.2 Description of Target Users

- **Key Demographics:**
  1] Age range: 12-45 years old
  2] Gender: Male/Female/others
  3] Job Titles: student, Researcher, Financier, Teacher, Software Engineer

- **Key psychographics**
  1] Explores more about particular topic
  2] Loves to learn new things

2.3 Advantages/Applications of Proposed System

- Detection of covid-19 will be in 5 seconds
- The proposed methodology is cost efficient and user friendly.

2.4 Scope

- The model is restricted to data samples only of cough and breath.

- The audio file should only be in the .wav format.

- The model is restricted to predict only Covid-19

SOFTWARE REQUIREMENT SPECIFICATION

3. Software Requirement Specification

3.1 Overview of SRS

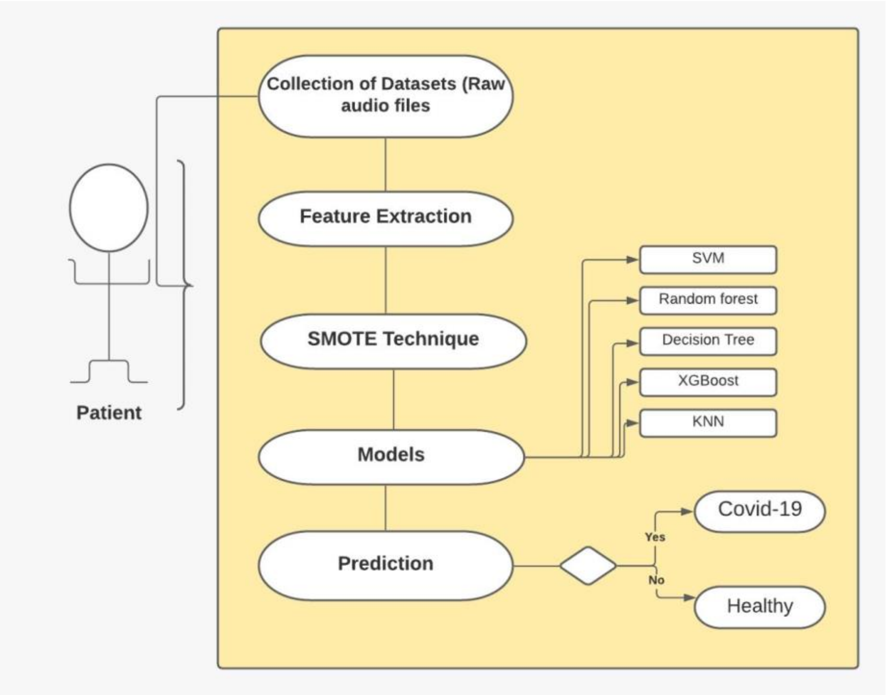
- Software should be maintainable, reliable, efficient, acceptable.
- Software should have aim to save life
- Software should be validated by the user.

3.2 Requirement Specification

3.2.1 Functional Requirements

- The model shall be able to read the cough and breath sound data.
- Model shall be able to classify various types of respiratory diseases.
- Model shall be able to analyze the given cough sound.
- The model should be able to display the predicted respiratory disease analysis.
- Developers should be able to log into the system remotely.
- Developers should be able to develop and improve the model.

3.2.2 Use case diagram



3.2.3 Use Case descriptions using scenarios

Use Case	Use case for question answering system
Actor	User
Precondition	Noise removal to be done for the recorded audio.
Goal	Respiratory disease classification
Scenario	Using machine learning concepts for prediction.
Exceptions	Only Covid-19 is able to detect.

3.2.4 Non-functional Requirements

3.2.4.1 Performance requirement

- The system should be able to classify and provide the analysis for respiratory disease within a span of120 seconds.
- The performance accuracy should be at least 90%.

3.2.4.3 System Requirements

- Linux and Python required for implementation.
- The data servers containing the data should be running 24/7 for clinical use.

3.3 Software and Hardware requirement specifications

3.3.1 Software Requirements

- Python

- Jupyter/ Google Collab

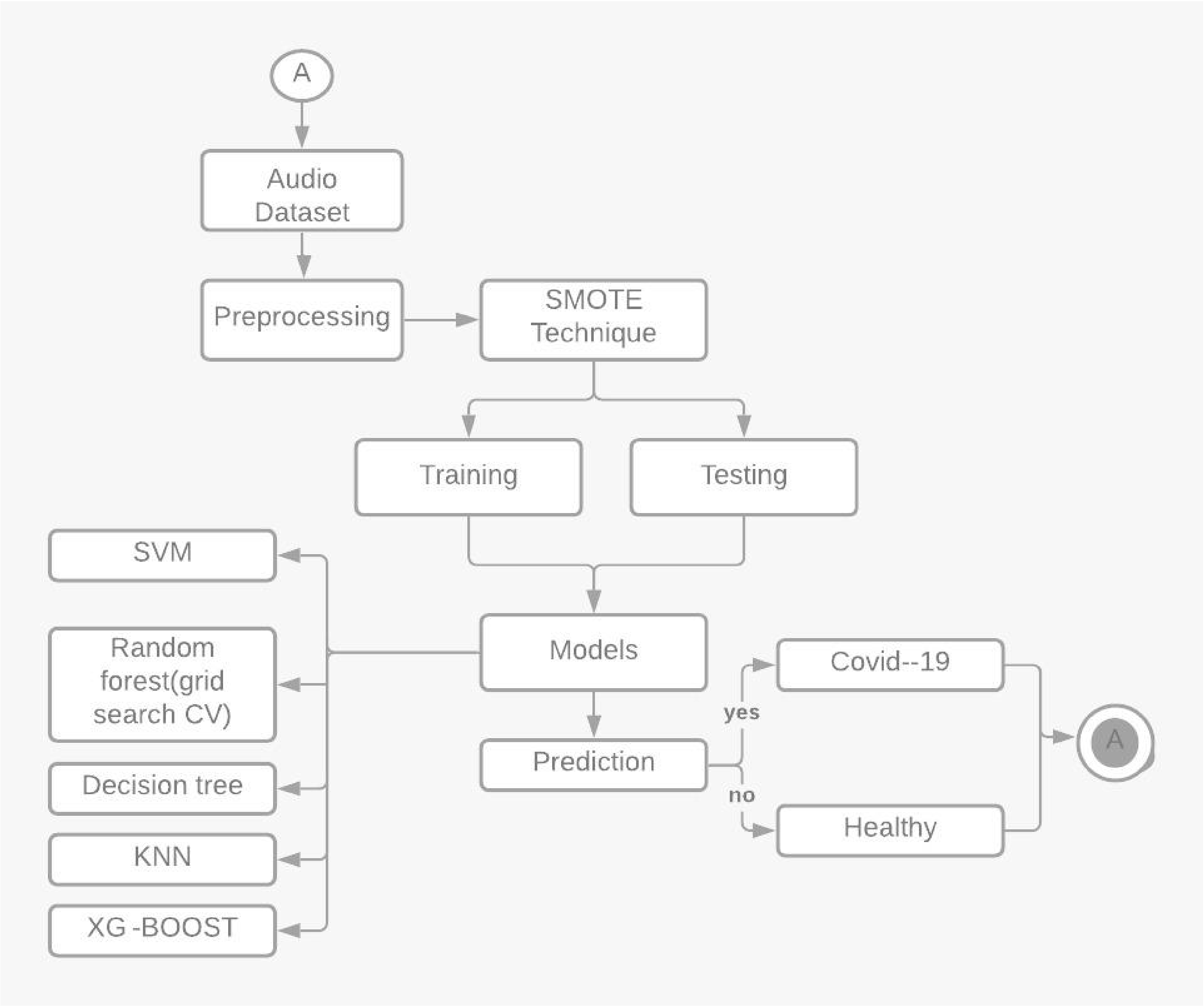
3.3.2 Hardware Requirements

- CPU: Intel i5 8th Generation (3.4 Giga Hertz)
- RAM = 16 GB
- Storage = 256 GB (SSD)

SYSTEM DESIGN

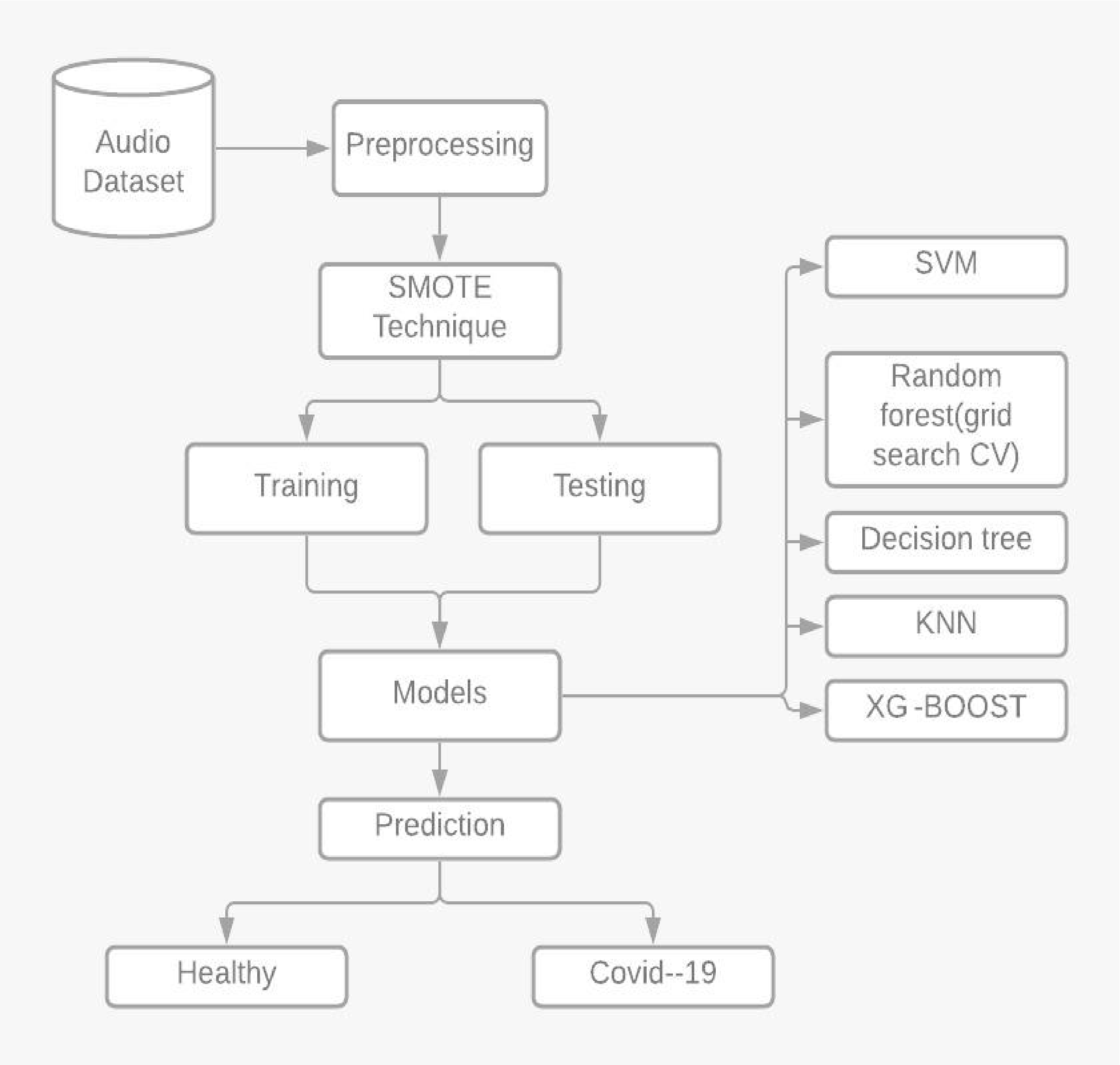
4. System Design

4.1 Architecture of the system



4.2 Architecture of the system

4.2.1 Diagram



5. Implementation

5.1- Data Description :

There are a total of 2 different datasets considered which are the Coswara dataset and the Coughvid dataset. These contain audio cough samples by the users recorded on microphone containing healthy and non healthy samples.
The Coswara dataset contains
The CoughVid dataset contains`

5.2- Methodology

SVM: Support Vector Machine is one of the most popular supervised learning algorithms used primarily for classification problems in the field of machine learning. The goal is that to create the decision boundary that can segregate n dimensional space into classes so that new data points can be put into the correct category for the future.
This decision boundary is called the hyperplane.

Random Forest With Grid Search: It is a supervised learning algorithm by constructing many decision trees at training level. The output obtained from this classifier is the category/class selected by most of the decision trees present in random forests.

KNN: K nearest neighbour assumes the similarities between new data and available cases. It stores all the available data and classifies new data points based on the similarity.

XGBoost: Extreme Gradient Boosting algorithm which is based on decision tree and uses boosting techniques to improve overall performance. Since it is having various frameworks, it is suitable for vast applications in data science challenges and classification problems.

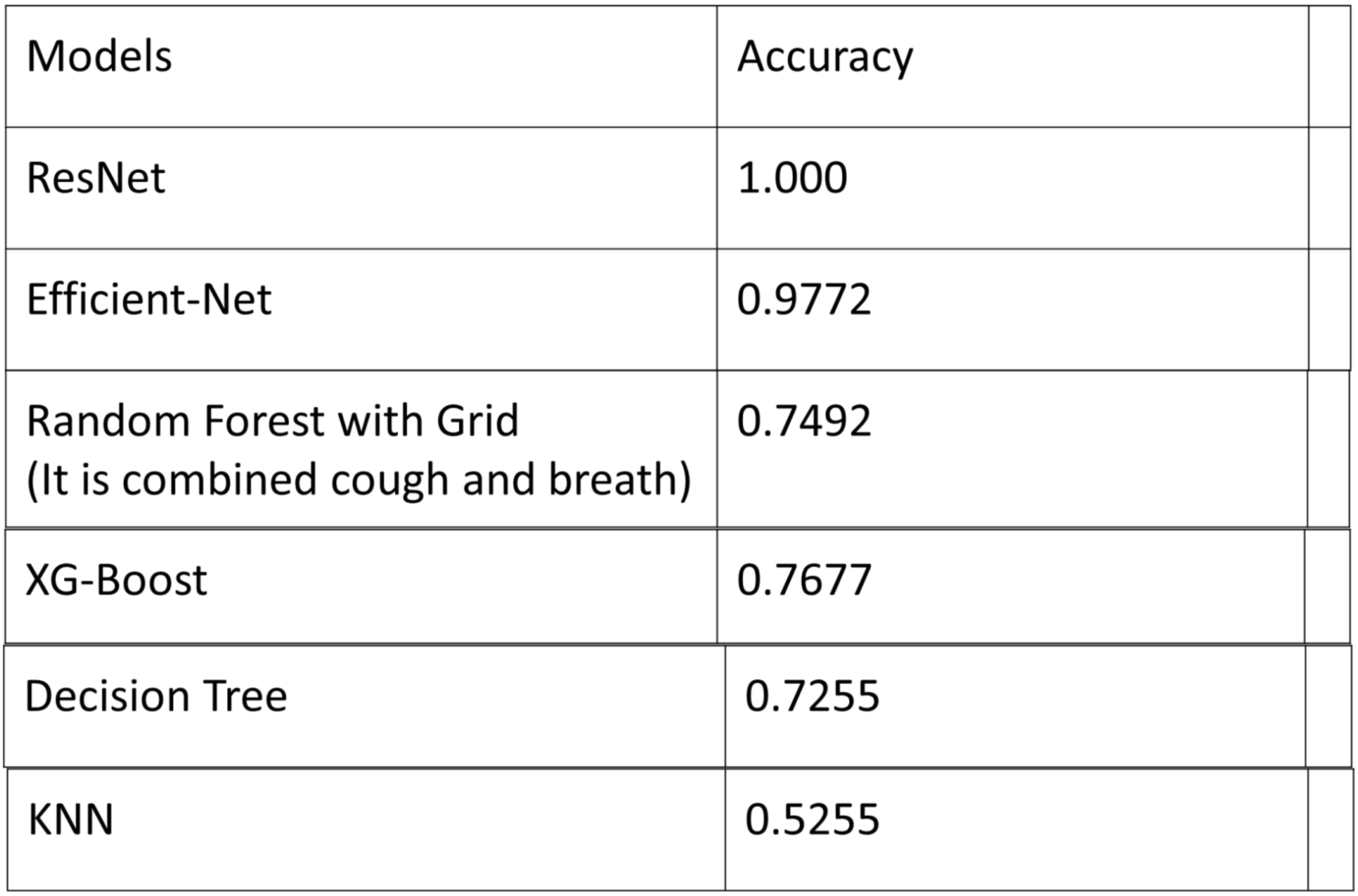
6. Testing

6.1 Acceptance test plan

Test id	Input	Expected output	Actual output
1	Requested to add testing file	Display a option add a file	Display a option add a file
2	Accepts the audio file (wav)	It should accept the file	Accepted the audio file (wav)
3	Duration of result output	Show the result within 1 min	Showed the result

Table 6.1.1- A

7. RESULTS & DISCUSSIONS



Models	Accuracy	
ResNet	1.000	
Efficient-Net	0.9772	
Random Forest with Grid (It is combined cough and breath)	0.7492	
XG-Boost	0.7677	
Decision Tree	0.7255	
KNN	0.5255	

These are following results we obtained

8. Conclusion and Future Scope

With the help of this approach, early detection of covid-19 in a much more accurate way would result in convenient testing. This would be also cost effective and safe as each covid positive user would record their samples remotely through their phones.

The future scope would be increasing the dataset and using this methodology for predicting covid 19 in different parts of the globe.

9. References:

[1] https://github.com/iiscleap/Coswara-Data for the Coswara dataset.

[2] https://github.com/virufy/virufy-covid for Virufy dataset from GitHub.

[3] https://zenodo.org/record/4048312#.YRtddXzivIU for Public Dataset CoughVid.

[4] J. Hada and Y. Takeuchi, "Evaluation of mental stress byvoice analysis during simulated landing, " The JapaneseJournal of Ergonomics, vol.44, No.3, pp. 171–174, 2008.

[5] J. Kaur and R. Kaur, "Extraction of Heart RateParameters Using Speech Analysis, " InternationalJournal of Science and Research, vol. 3 issue 10, pp.1374–1376, 2014.

[6] Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, Handling imbalanced datasets: areview. GESTS International Transactions on Computer Science and Engineering,2006. Vol 30(No 1): p. 25-36.

[7] Zhu, Z.-B. and Z.-H. Song, Fault diagnosis based on imbalance modified kernelFisher discriminant analysis. Chemical Engineering Research and Design, 2010.88(8): p. 936-951.

[8] Sujatha, C.M., Mahesh, V., Ramakrishnan, S.: Comparison of two ANN methods for classification of spirometer data. Measurement Science Review 8(3), 53–57 (2008)

[9] Jude, H.D., Kezi, S.V.C., Anitha, J.: Application of Neuro-Fuzzy Model for MR Brain Tumor Image Classification. Biomedical Soft Computing and Human Sciences 16(1), 95–102 (2009)

[10] Uncu, U.: Evaluation of Pulmonary Function Tests by Using Fuzzy Logic Theory. Journal of Medical Systems 34(3), 241–250 (2010)

BIBLIOGRAPHY

Lara Orlandic September 2020

- The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms

- https://www.researchgate.net/publication/344372987_The_COUGHVID_crowdsourcing_dataset_A_corpus_for_the_study_of_large-scale_cough_analysis_algorithms