KLE Society's
KLE Technological University, Hubballi.

A Minor Project Report

On

COVID-19 DISEASE DETECTION

submitted in partial fulfillment of the requirement for the degree of

Bachelor of Engineering

In

School of Computer Science and Engineering

Submitted By

Shreyas N B	01FE19BCS005
Ramakrishna MutalikDesai	01FE19BCS024
Sharath S Shanbhag	01FE19BCS029
Nikhil Kurane	01FE20BCS423

Under the guidance of
Prof. Umadevi F.M.

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBBALLI – 580 031

Academic year 2021-22

KLE Society's
KLE Technological University, Hubballi.

2021 - 2022

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Minor Project entitled COVID-19 Disease Detection is a bonafide work carried out by the student team Shreyas N.B. USN:01FE19BCS005, Ramakrishna MutalikDesai USN:01FE19BCS024, Sharath S Shanbhag USN:01FE19BCS029, Nikhil Kurane USN:01FE20BCS423 in partial fulfillment of completion of Sixth semester B.
E. in Computer Science and Engineering during the year 2021-2022. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the above-said program.

Guide

Prof.Umadevi F.M.

Head, SoCSE

Dr. Meena S. M

External Viva -Voce:

Name of the Examiners	Signature with date
1.	
2.	

Acknowledgement

We would like to thank our faculty and management for their professional guidance towards the completion of the project work. We take this opportunity to thank Dr. Ashok Shettar, Vice-Chancellor, Dr. N.H Ayachit, Registrar, and Dr. P.G Tewari, Dean Academics, KLE Technological University, Hubballi, for their vision and support.

We also take this opportunity to thank Dr. Meena S. M, Professor and Head, SoCSE for having provided us direction and facilitated for enhancement of skills and academic growth.

We thank our guide Ms.Umadevi F.M., Professor, SoCSE for the constant guidance during interaction and reviews.

We extend our acknowledgement to the reviewers for critical suggestions and inputs. We also thank Project Co-Ordinator Mr. Uday N.Kulkarni, Mr. Guruprasad Konnurmath and all the reviewers for their support during the reviews and course of completion.

We express gratitude to our beloved parents for constant encouragement and support.

Shreyas N B - 01FE19BCS005
Ramakrishna MutalikDesai - 01FE19BCS024
Sharath S Shanbhag - 01FE19BCS029
Nikhil Kurane - 01FE20BCS423

ABSTRACT

In late 2019, coronavirus-2, a significant human virus that caused the blockage of airways, lung regions, pulmonary vascular tubes, and respiratory neuro-muscular disorders.
The standard procedure to detect this virus is RT-PCR, which has a higher success rate,but it has a numerous drawbacks which includes expensive equipments and requirement of chemical agents, the need for expertise of the nurses and doctors for diagnosis,and it takes a lot of time to process the diagnosis. To overcome this, A deep learning model would guarantee elimination of these drawbacks. During the COVID-19 pandemic, it can be considered as an early, quick time-efficient, and nearly cost-free diagnosis tool that complies with social distance constraints. This projects aims to endeavour over the limitations of the current standard and provide a beneficial solution to the common public since it reduces the work of the doctors and the time constraint is also resolved.

Keywords :*Respiratory Disease, coronavirus-2, RT-PCR anti-body tests.*

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

COVID-19 has created a havoc in all the regions of the world, halting life and claiming thousands of lives all across the world. COVID-19 has spread to many nations and territories, posing a major threat to public health, with a growing number of infected patients and death rate tolls .

Coronavirus is a common virus that causes infections in the nose and throat region. COVID-19, according to the researchers, is a sickness caused by the SARS-CoV-2 virus that can causes blockage of airways ,difficulty in breathing, and with various other symptoms.The clinical diagnosis procedure today necessitates in-person involvement, demands flexibility , and is expensive. It is not even available to the general population in several poor nations due to a paucity of healthcare facilities. As a result, a low-cost, quick, and easily accessible approach to identifying COVID-19 is a requirement in most parts of the country.

Despite this, a number of research projects evaluating potential cures and treatments are now ongoing. Researchers have begun to employ a variety of machine learning and deep learning techniques and approaches to identify Coronavirus with the use of images of X-rays and CT scans , extracting appropriate features . The most important stumbling problem is that just a few datasets are available. Furthermore, fewer specialists are available to keep track of the facts around the rising amount of pollution causing diseases in humans.

1.1 Motivation

The continuation of this epidemic is leading to an increasing number of confirmed cases and people suffering from severe internal organ failures, There are good reasons to worry about the eradication of this deadly viral disease. The adoption of a number of this virus-related difficulty-solving strategies has aroused much interest. However, another important issue researchers need to overcome is the growing volumes of data, which presents challenges in this battle against the virus.This shows how Machine Learning / In-Depth Learning can be very important in development of health care systems around the world.

These emerging techniques have save a lot of time in a period when even an hour of time saved could result in the saving of lives in almost all possible locations where Coronavirus is spreading among the humans .With the growing demand of Deep Learning applications in clinical contexts, it can play an important role in improving productivity and efficiency in studies involving great increase in amounts of data and requiring increasing levels of accuracy in detection and diagnosis.

1.2 Literature Review / Survey

1.2.1 Detection of COVID-19 in smartphone-based breathing recordings

The aim of this study was to scrutinize whether smartphone-based methods can be used to classify between COVID -19, including asymptomatic and healthy individuals, within a deep learning framework. The publicly available Coswara database, contained a total of 480 breath sounds (240 deep and 240 shallow). A deep learning framework is proposed based on manually generated features extracted from the original recordings and Mel frequency cepstral coefficients , and deep activated features learned using a combination of CNN and bidirectional short-term memory units.

The proposed deep learning technique showed a performance accuracy of 0.9458 and 0.9208 when using shallow and deep images, respectively. Most importantly, it effectively detected COVID -19 individuals with a maximum sensitivity of 0.9421, specificity of 0.9496, and AUC-ROC of 0.90. symptom less individuals were detected with an accuracy of 0.1 on shallow recordings and 0.8889 on deep recordings.

Key points:

- Mohanad Alkhodari, Ahsan H. Khandoker; January 13,2022
- Method/Model:Feature extraction:Combining of hand-crafted features and deep activated features by the combination of CNN and Bi-LSTM.
- Dataset:Coswara :breathing samples
- Performance:Accuracy: [breathing-shallow:0.9458], [breathing-deep:0.9208]

The gaps identified/ technical challenges :

- The collection solely contains breathing samples, but the entire Coswara dataset contains breath, cough, and voice sounds.

1.2.2 Rapid and Scalable COVID-19 Screening using Speech, Breath, and Cough Recordings.

This article aims to provide the reader a sense of classifying audio samples into positive and negative samples. Lately, The technologies offer greater access and more frequent and rapid testing, but comes with drawbacks, including the robustness of the procedures and the hygienic nature of forced cough recordings. In this study, we present a new technique for grading patients based on acoustic features in speech and breath sounds. Mel frequency cepstral coefficients–MFCC and relative spectra perceptual linear prediction–RASTA-PLP features are evaluated individually and combining two algorithms, Random Forests (RF) and Deep Neural Networks (DNN). With an AUC score of 0.7938 for identifying COVID -19 via analysis of speech sounds and 0.7575 for identifying COVID -19 via analysis of breath sounds, the combination of MFCC and RASTA-PLP provides the best results for speech and breath sounds.

Key points:

- Published : Drew Grant, Ian McLane, and James West,2021 IEEE EMBS.
- Method/Model:Deep Neural Network,Random Forest.
- Dataset:Coswara Dataset[Cough, Speech and Breath sound recordings.]
- Performance:[Cough:DNN:0.68 ,RF:0.66],[Speech:DNN:0.75, RF:0.71],[Breath:DNN:0.73, RF:0.67].

The gaps identified/ technical challenges :

- The labels here are relying of self-assessment of the users who are volunteering their information.

1.2.3 Automated detection of COVID-19 cough.

This article talks about the possibilities in automatic detection of this deadly virus,the COVID -19 from cough. Fitting a RF model with the set of time-frequency features selected by RFE for discriminating COVID -19 cough gave excellent results. This new method could lead to automatic COVID -19 identification with existing simple and portable devices. We will collect more high-quality data, especially different cough patterns from other diseases, and curate the current corpus to train, tune, and improve the performance of the models. Other previous research extracted MFCCs from cough recordings and fed them into a pre-trained CNN. Their model had an AUC of 0.97 and a specificity of 0.942.A machine learning-based

COVID -19 cough classifier capable of distinguishing COVID -19 positive cough from COVID -19 negative and healthy cough recorded with a smartphone was presented. Using the Resnet50 classifier to discriminate between COVID -19 positive and healthy cough, they achieved an AUC of 0.98, and using the LSTM classifier to discriminate between COVID -19 positive and COVID -19 negative cough, they achieved an AUC of 0.94. In all tasks, their models had an AUC greater than 0.80.

Key points:

- Published :13 September 2021
- Method/Model:-YAMNet Architecture,Auto-Encoder Architecture,Recursive Feature Elimination.
- Dataset:University of Cambridge: Coswara Virufy.
- Performance:Random Forest performed better[Accuracy:0.8367, Sensitivity:0.8958, Specificity:0.7158, Precision:0.8662, F-Score:0.8804, AUC:0.9356]

The gaps identified/ technical challenges :

- Correctly classifying negative samples is an issue and to make data corpus curate the comparing COVID-19 cough patterns with cough patterns from other conditions, such as asthma or bronchitis, are needed.

1.2.4 Pay attention to the speech: COVID-19 diagnosis using machine learning and crowdsourced respiratory and speech recordings.

Since the havoc created by the deadly virus,the covid-19, many efforts have been put to use the breath audio and cough samples collected by smartphones to distinguish between healthy and covid-19. With the help of python libraries ,these models can be transformed into mobile apps with IoT connectivity features.Using specific voice patterns in addition to other breathing sounds can improve the results. A total of 9 different types of sounds were recorded, which were labelled with the status COVID -19. A combination of models trained on different sounds can diagnose COVID -19 with more accuracy than a single model trained only on coughing or breathing. The results show that by taking the average of the predictions of multiple models trained and evaluated separately on different sounds, an AUC of 0.964 and an accuracy of 0.96 with simple binary classifiers can be achieved.

Key points:

- Published:23 August 2021
- Method/Model:Feature Extraction:MFCC,Spectral,centroid,Spectral,bandwidth,Roll-off and ZCR[CNN architecture].
- Dataset:Coswara Dataset
- Performance:[Breathing Sounds:Accuracy:0.81],[Cough sound:Accuracy:0.84],[Speech Sound:Accurac 0.71],[vowels sound:AUC:0.96]

The gaps identified/ technical challenges:

- Coswara Dataset was imbalanced but with SMOTE technique,this issue was solved.

1.2.5 An ensemble learning approach to digital corona virus preliminary screening from cough sounds.

The cough sounds recorded contains a variable amount of cough samples. Accurate detection of COVID -19 from the sound data sets requires overcoming two major challenges: (i) the varying number of coughs in each recording and (ii) the small amount of COVID -19 positive cases compared to healthy cough samples in the dataset. A total of two crowd sourced open datasets of cough records from the population and each cough record is segmented into non-overlapping coughs. This segmentation enhances the original data without oversampling by splitting the original cough sound into non-overlapping segments. Splitting the cough sounds allows us to increase the number of minority class samples which is COVID -19 , without changing the feature distribution of the COVID -19 samples created by using oversampling techniques. Each cough sound segment is converted into six image representations for further analysis. We perform experiments using Convolutional Neural Network , and pre-trained CNN models. The results of our models were compared with other recently published machine learning works with cough sound data to detect COVID -19.

Key points:

- Published:28 July 2021
- Authors: Emad A. Mohammed, Mohammad Keyhani, Amir Sanati-Nezhad, S. Hossein Hejazi Behrouz H. Far .
- Method/Model:Feature Extraction:spectrogram, power spectrum, MFCC, and Mel-Spectrum.[VGG-16 architecture].

- Dataset:Coswara Dataset(cough and breathing sounds)
- Performance:[Random Forest (RF)-accuracy :0.67]

The gaps identified/ technical challenges:

- To bring the difference among class labels into limelight and minimize excessive false positive or false negative results.

1.3 Problem Statement

Detecting Covid-19 from Audio samples using Deep Learning Techniques.

1.4 Applications

- Medical Imaging and Diagnostics
- Simplifying Clinical Trials
- Personalized Treatment

1.5 Objectives and Scope of the project

The objective determines the output for a desired operation. Scope determines the applicability of the project in the real world.

1.5.1 Objectives

- Identification of appropriate audio features such as MFCCs, Spectral centroid, RMS and ZCR.
- Applying Machine Learning algorithms like Random Forest, KNN and SVM.
- To classify the audio samples into COVID-19 and healthy.
- Performance evaluation of the deep learning models against machine learning models.
- Comparing with state of the art works.

1.5.2 Scope of the Project

- The model is restricted to data samples only of cough and breath audio samples.
- The model restricted to predict only Covid-19.
- Detection of Covid-19 will be within the span of 120 seconds.
- With deep learning techniques, patient care details can be delivered by analyzing patients' medical history along with symptoms.

Chapter 2

REQUIREMENT ANALYSIS

It is a procedure of measuring the requirements and expectations of a new product and there should be often communications among the stake-holders and end-users of the product to acquire their expectations, resolving the existing conflicts, and document all the important requirements. The biggest challenge one can find is to provide the visualization of the final products with the clients.

2.1 Functional Requirements

Functional requirements defines about a product's future usage.

- Read the cough and breath audio data.
- To classify audio sound into healthy and covid-19 diseases.
- Model shall be able to analyze the given cough sound.
- The model should be able to display the predicted respiratory disease analysis.

2.2 Non Functional Requirements

Non-functional requirements gives a description about how well the systems can be maintained and scalable over various devices and architectures.

- The performance accuracy should be at least 0.90 .
- The system should be able to classify the given audio into COVID-19 as well as healthy within a span of 120 seconds.

2.3 Hardware Requirements

- RAM - A minimum of 8 Gigabytes.
- Processor - Intel I5/AMD Ryzen 5 or higher.

- GPU - A minimum of 4 Gigabytes.
- Graphics - NVIDIA Ge-force /AMD Radeon

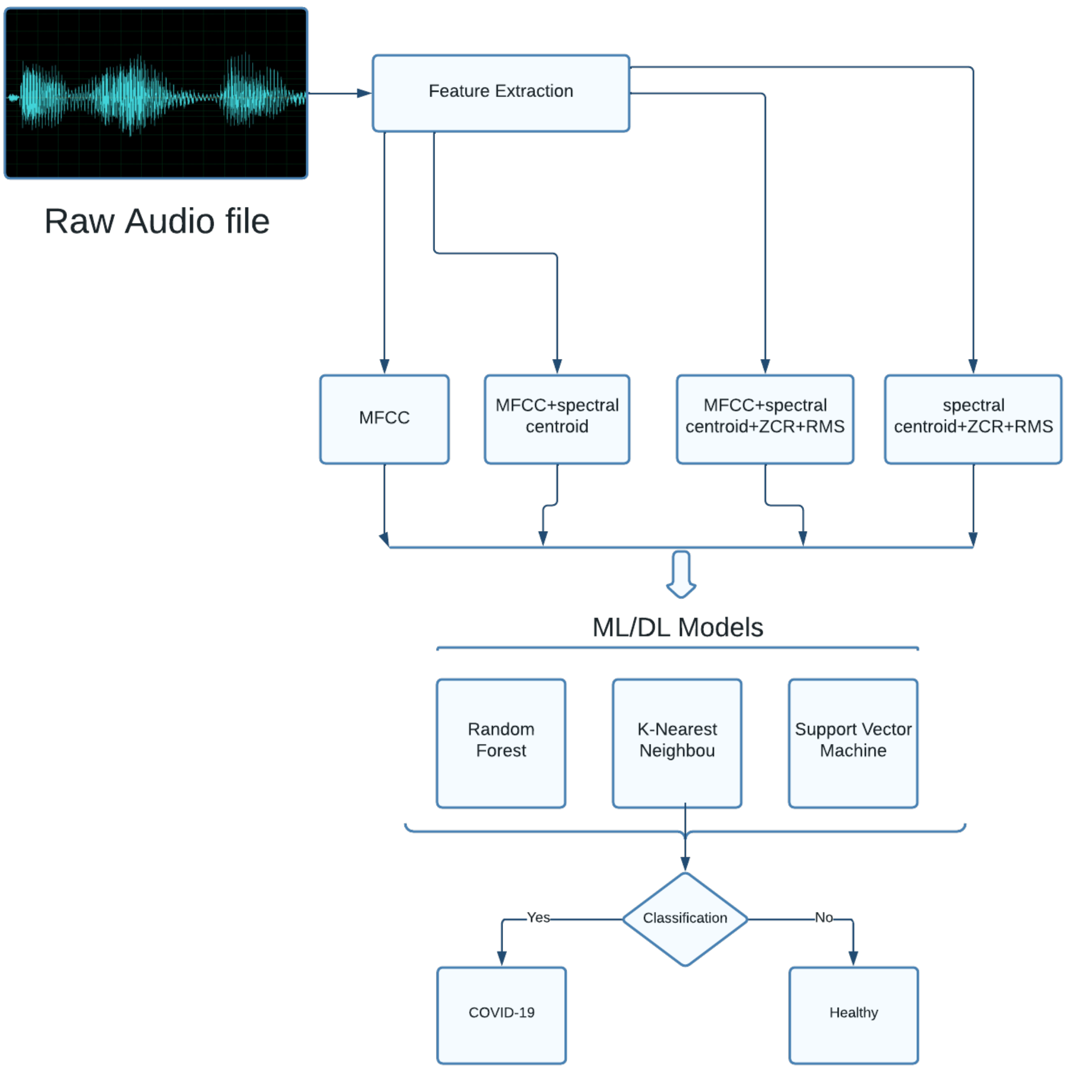
2.4 Software Requirements

- Google Colab - To compile and run Notebooks.
- Google Drive - To store and retrieve the datasets.
- Python - Programming language.

Chapter 3

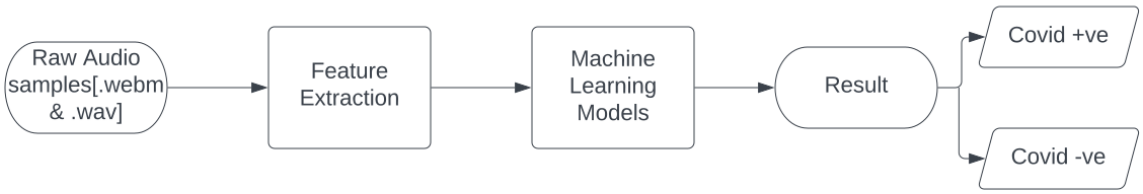
SYSTEM DESIGN

System design is an engineering discipline that allows the creation of successful systems.The system design process specifies software and hardware components, linkages, and data in order for the system to meet a set of well-defined key requirements.

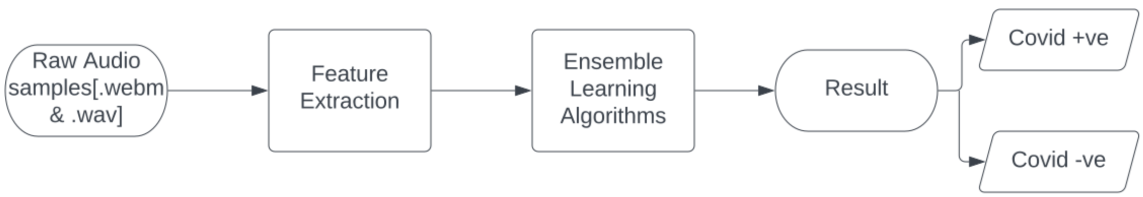


3.1 Architecture Design

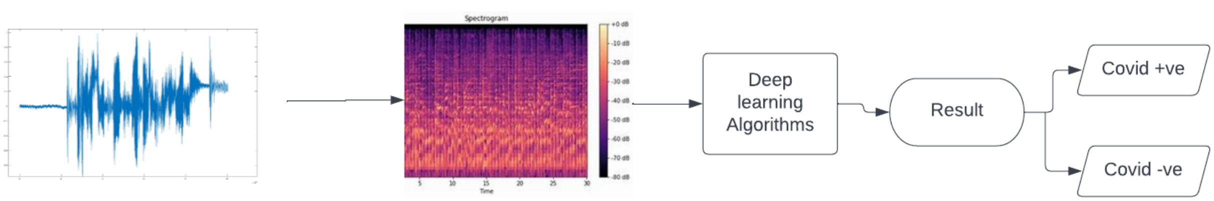
1.Machine Learning : The most applicable audio features, such as MFCC, Spectral centroid, RMS, and ZCR, are retrieved from the raw audio samples from the datasets. With the use of python libraries, these collected features are fed into various machine learning models for the detection of covid-19.



2.Ensemble Methods : For the detection of covid-19, the most appropriate audio features are extracted, and fed into an ensemble of SVM, Logistic regression, and Decision tree layers.



3.Deep Learning : Because the mel spectrogram converts hertz data to mel scale values, the raw audio files are converted to mel-spectrograms. The mel spectrograms are better for situations where human hearing perception needs to be simulated. Deep learning methods are then used to detect covid-19 using these spectrogram images .



3.2 Feature Engineering

When creating a prediction model with machine learning, feature engineering is the process of gathering domain expertise in order extract the most essential characteristics from raw audio data.

The common features used are :

- **MFCC :**MFCCs is a type of characteristics that define the overall shape of a spectral feature of the audio sample.
- **Spectral centroid:**Spectral centroid tells about the location of center of mass of the spectrum .
- **ZCR :**It is the rate at which , the given signal transitions from positive to zero to negative or from negative to zero to positive.
- **Root Mean Square** :When using RMS with audio, the signal values (amplitude) are squared and averaged over time to determine the square root of the result.

Table 3.1: Feature Combinations

Feature	**Feature Combinations**
F1	MFCC
F2	MFCC + Spectral centroid
F3	MFCC + Spectral centroid + RMS + ZCR
F4	Spectral centroid + RMS + ZCR

Chapter 4

IMPLEMENTATION

4.1 MACHINE LEARNING

The machine learning models used namely KNN(K-Nearest Neighbor),Random Forest, and SVM(Support-Vector-Machine).

1. KNN: The suggested KNN model divides the input signal into two categories namely healthy and Covid-19. Predicts the values based on similarities in features, which implying that the new data points will be assigned a value depending on how closely it matches the points in the training set. With the help of the steps below, we can grasp how it works.

- Step 1 - The dataset provided is split into train and test datasets.Then ,both training and test data is loaded.
- Step 2 Next, the value of K must be selected where k is the closest data points and it can be any positive integer.
- Step 3 Perform the following for each point in the test data:
- 3.1 Determine the distance between the test data and each row of training data using any of the following methods: Euclidean, Manhattan, or Hamming distance. The most often used method is the Euclidean distance for calculating distance.
- 3.2 Arrange them in ascending order depending on the distance value.
- 3.3 It will then choose the top K rows from the sorted array.
- 3.4 It will now assign a class to the test point based on the most often occurring class in these rows.
- Step 4 finish.

2. Random Forest : It is a classification technique that uses many decision trees to classify data. It use bagging and feature randomization to generate an uncorrelated forest of trees whose overall prediction is more accurate than that of any single tree when each individual tree is constructed.

- Step 1 Begin by randomly selecting samples from a specified dataset.
- Step 2 Following that, this algorithm will create a decision tree for each sample. The forecast result from each decision tree will then be obtained.
- Step 3 Voting will be done for each expected outcome in this stage.
- Step 4 Finally, choose the most popular forecast result as the final prediction result.

3. SVM : SVM is one of the accurate classification methods available. Converts the inputs into its feature vectors, the kernel technique enables SVMs to perform non-linear classification effectively.

The fundamental goal of SVM is to divide datasets into classes which may be done in two steps:

- Initially, The SVM algorithm will generate hyper-planes that best separate the classes repeatedly.
- The hyperplane that best distinguishes the classes will be chosen subsequently.

4.2 ENSEMBLE METHODS

A machine learning ensemble is a model that combines the predictions of two or more models.It is stated that this combination of different layers from the models can perform well than a single model.

The ensembled model used includes SVM, logistic regression, and Decision tree.
SVM : is based on the idea of getting the largest distance between the points of the dataset and the separation hyperplane.
Logistic Regression : It is a framework for statistical learning that combines principles from regularisation as well as ensembling. This framework is used to create a binary classification ensemble of logistic regression models.
Decision tree : Applying a certain threshold to a specific feature and splitting the points into two groups, a large number of points may be assigned to a single class.

4.3 DEEP LEARNING

ResNet : ResNet is a deep learning model that is applicable in a wide range of computer vision tasks. It employs skip connections to add the output from a previous layer to a current layer .

DenseNet : It is a type of CNN that results in dense connections between layers. It follows feed-forward nature as each layer takes inputs from all the previous layers and passes on ,its own feature-maps to the next layers.

Chapter 5

RESULTS AND DISCUSSIONS

5.1 Datasets

5.1.1 COUGHVID Dataset

- This dataset contains information about a wide range of subjects' ages, genders, geographical loactions, and Corona-virus statuses.
- There are a total of 19213 segmented cough audio samples in .webm and .ogg formats.

5.1.2 COSWARA Dataset

- This dataset includes breathing sounds, cough sounds , and vowels .
- Breathing-deep audio samples has a total of 746(542 healthy and 204 covid positive samples).
- Cough-heavy audio samples has a total of 765(574 healthy and 191 covid positive audio samples).
- Vowels audio samples has a total of 745 with 552 healthy and 193 covid positive voice samples.

5.2 Performance Evaluation

The performance metrics such as precision, accuracy, F1 score and AUC scores along with their respective values for each model is shown in the tables below and visualizations such as confusion matrices as well as train vs validation curve graphs are plotted.

5.2.1 MACHINE LEARNING

Table 5.1: Machine Learning-Coughvid Results

Models	Feature Combinations	Precision	Recall	F1 score	AUC
KNN	F1	0.9200	0.9064	0.9047	0.9064
	F2	0.8922	0.8771	0.8745	0.8771
	F3	0.8922	0.8771	0.8745	0.8771
	F4	0.8922	0.8771	0.8745	0.8771
Random Forest	F1	0.9923	0.9922	0.9922	0.9928
	F2	0.9961	0.9961	0.9961	0.9964
	F3	0.9961	0.9961	0.9961	0.9964
	F4	0.9961	0.9961	0.9961	0.9964
SVM	F1	0.5178	0.5107	0.5114	0.5886
	F2	0.6258	0.6151	0.6114	0.6192
	F3	0.6179	0.6107	0.6086	0.6107
	F4	0.6258	0.6151	0.6114	0.6192

Table 5.2: Machine Learning-Coswara Results

Models	Feature Combinations	Precision	Recall	F1 score	AUC
KNN	F1	0.625	0.65	0.6464	0.65
	F2	0.7564	0.725	0.7163	0.725
	F3	0.7564	0.725	0.7163	0.725
	F4	0.7564	0.725	0.7163	0.725
Random Forest	F1	0.7604	0.75	0.7474	0.75
	F2	0.7020	0.7	0.6992	0.7
	F3	0.6790	0.675	0.6731	0.675
	F4	0.7020	0.7	0.6992	0.7
SVM	F1	0.578	0.75	0.55	0.6307
	F2	0.5789	0.5796	0.5791	0.5756
	F3	0.5789	0.5796	0.5791	0.5756
	F4	0.5791	0.5796	0.5791	0.5756

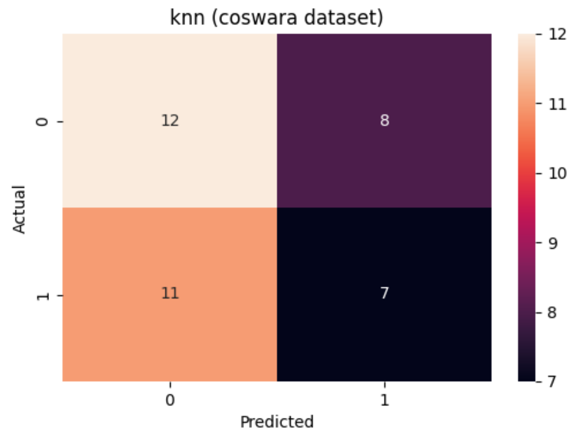


Figure 5.1: KNN-Coswara

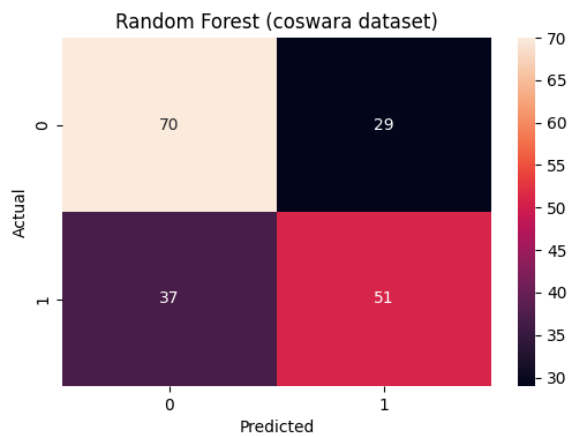


Figure 5.2: RandomForest-Coswara

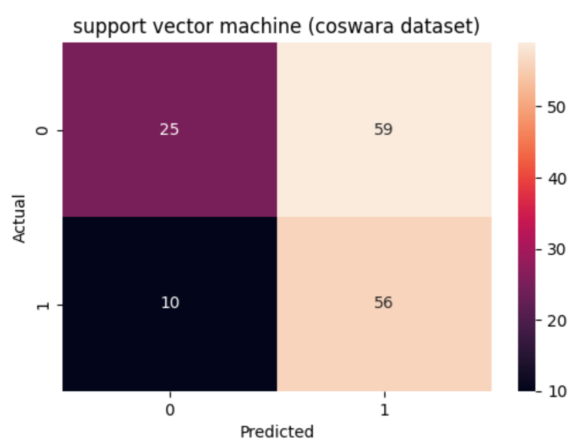


Figure 5.3: SVM-Coswara

5.2.2 ENSEMBLE METHODS

Table 5.3: Ensembled Model-Coughvid Results

Models	**Feature Combinations**	**Precision**	**Recall**	**F1 score**	**AUC**
SVM + Logistic regression + Decision tree	F1	0.8048	0.7767	0.7905	0.7942
	F2	0.6787	0.8438	0.7523	0.722
	F3	0.6779	0.8434	0.7517	0.7214
	F4	0.6060	0.8077	0.692	0.6412

Table 5.4: Ensembled Model-Coswara Results

Models	**Feature Combinations**	**Precision**	**Recall**	**F1 score**	**AUC**
SVM + Logistic regression + Decision tree	F1	0.8127	0.8028	0.8028	0.9005
	F2	0.8487	0.8339	0.8339	0.8856
	F3	0.8303	0.8180	0.8180	0.8930
	F4	0.8265	0.8145	0.8145	0.8939

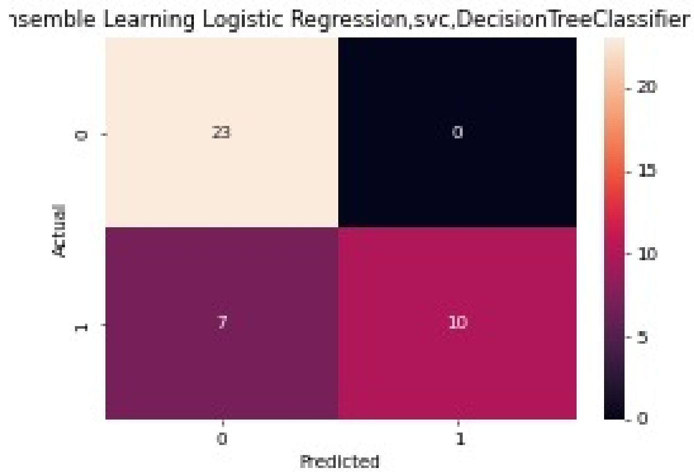


Figure 5.4: Ensemble Model+F1

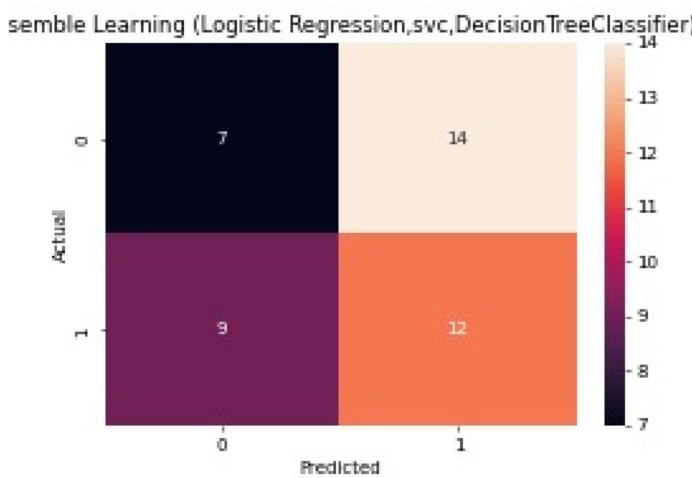


Figure 5.5: Ensemble Model+F2

5.2.3 DEEP LEARNING

Table 5.5: Deep Learning Results

Models	**Accuracy**
ResNet	0.8133
DenseNet	0.9083

Figure 5.6: ResNet-TrainVsValidation Graph

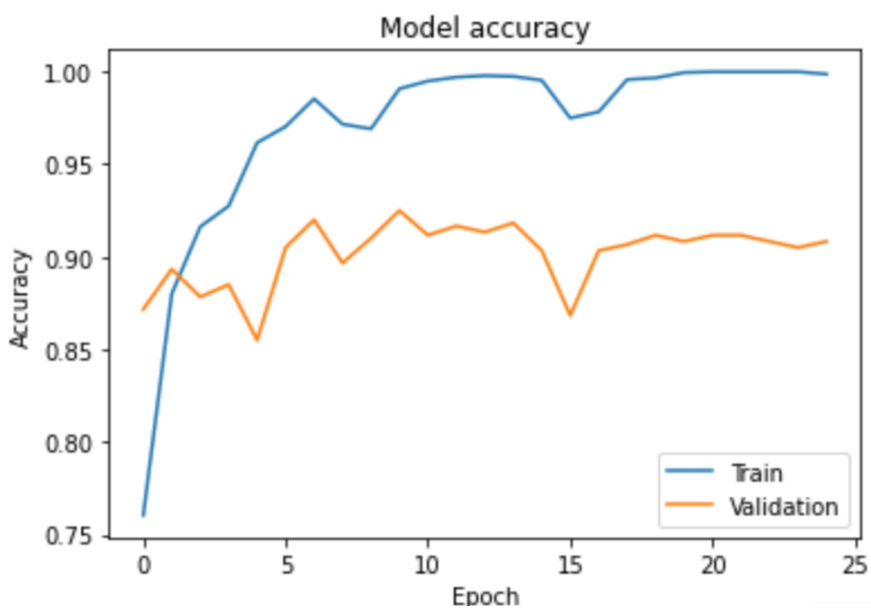


Figure 5.7: DenseNet-TrainVsValidation Graph

Chapter 6

CONCLUSIONS AND FUTURE SCOPE

6.1 Conclusion

To increase the prediction performance, such as higher classification accuracy, we explicitly use ensembled model. Deep learning relies on iterative learning methods that exposes the machines to large amounts of data and makes it extremely powerful when dealing with unstructured data .

Ensemble learning and Deep learning approaches have performed better than Machine learning techniques due to a lack of domain understanding for feature introspection. Since feature engineering is less of a worry in case of Deep Learning techniques outperform others.

6.2 Future Scope

The future works for this field would be to consider large number of samples of other classes of respiratory diseases as very less number of other samples are available currently which will lead to under-fitting model.