Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
Gazipur, Bangladesh

Chronic Diseases Prediction on COVID-19 Patients Using Machine Learning Techniques

by

Abeed Hanif Aurko (160041012)
S. M. Fahim Abid (160041020)
Dewan Tarikul Mannan (160041027)

Supervisor
Md. Hamjajul Ashmafee
Lecturer
Dept. of CSE, IUT

A thesis submitted in partial fulfilment of the requirements
for the degree of B. Sc. Engineering in Computer Science and Engineering

Academic Year: 2019-2020

March, 2021

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by under the supervision of Hamjajul Ashmafee, Lecturer of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Abeed Hanif Aurko
Student ID - 160041012

S. M. Fahim Abid
Student ID - 160041020

Dewan Tarikul Mannan
Student ID - 160041027

Approved By:

Md. Hamjajul Ashmafee
Lecturer
CSE, IUT

Prof. Dr. Abu Raihan Mostofa Kamal
Head of the Department
CSE, IUT

Acknowledgement

We would like to express our grateful appreciation for **Hamjajul Ashmafee**, Lecturer, Department of Computer Science & Engineering, IUT for being our adviser and mentor. His motivation, suggestions and insights for this research have been invaluable. Without his support and proper guidance this research would never have been possible. His valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way. We are really grateful to him.

We are also grateful to **Dr. Abu Raihan Mostofa Kamal**, Professor, Department of Computer Science & Engineering, IUT for his valuable inspection and suggestions on our proposal of Chronic diseases prediction on COVID-19 patients using machine learning techniques

We would like to also mention few expert with clinical knowledge for supporting us by evaluating our clinical information. Our heartist thanks to **Dr. Shahida Begum** FCPS, MCPS, DGO (Gynae / Obst) Consultant, Lab Aid, Barishal, **Dr. Jalal Uddin** Diploma in Anesthesia and ICU Consultant Anesthesiologist, Barishal and **Dr. Tanjima Kulsum** MBBS, Shahid Suhrawardy Medical College for their constant opinions and support while it came to medical information.

Abstract

COVID - 19 pandemic has spread to more than 210 countries. Millions of people lost their lives due to COVID - 19. COVID-19 death rate is almost 2.2%. So, the majority of people are recovering from the disease. But recently a lot of people recovered from COVID 19 are developing chronic diseases (ie. Heart failure, Stroke, Chronic Kidney disease, Liver damage, Chronic Obstructive Pulmonary disease, Shock, Blood Clotting etc). which is really alarming. In our work we tried to develop a combine system to predict the after COVID - 19 chronic disease probability. Here we first developed a central model which works well for different individual disease predictions (heart, lungs, kidney, liver diseases). Then we worked with COVID - 19 patients data to predict the probability of chronic diseases based on the changes in different haematological parameters. It will help the COVID - 19 recovered patients to take precautionary measures against chronic diseases to minimize the diseases and avoid the casualties.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Chronic Disease prediction on covid-19 patients

COVID - 19 pandemic [11] which is still ongoing costs millions of lives. More than 97.5% recovered from COVID - 19. But alarming news is most of them are developing a lot of after COVID complications [4] [22] as well as a lot of chronic disease [3]. Some post COVID complications are Hyperoxia (due to lack of oxygen), fatigue, severe organ damage, blood clotting problems. Recent study shows that 73% COVID survivors people suffer from mental illness and 1 out of 4 COVID recovered person (25%) are developing chronic diseases [3] like cardiac complications (due to blood clotting and blood vessels problems which occurs after COVID 19 recovery), liver damage, chronic kidney disease (CKD), Chronic obstructive pulmonary disease (COPD) and other chronic diseases. We developed a model to predict heart disease, kidney disease, liver disease, pulmonary diseases from different datasets. And we passed some COVID - 19 patients data through the model to get a prediction about the probability of chronic diseases.

1.2 Problem Statement

Main goal of this thesis is to build a model which will help the medical personals and researchers to predict the risk of chronic disease of COVID - 19 recovered patients. As the research is still going on to clinically identify this risk, our model will help them to predict the risk more sophisticatedly. Then people can take precautionary measures and casualties could be minimized.

There has been many theoretical finding to support the fact that COVID-19 increases the possibility of chronic diseases. But none is there to mathematically or practically proven. our problem is to address that.

1.3 Challenges working with big data in healthcare system

1.3.1 Lack of Organized health datas

It is very tough to find data organized in a way that can be served for predicting different cases. Most of the dataset are found in very different format.

1.3.2 Confusing variety of big data technologies

It is very tough to choose the perfect technologies from the market

1.3.3 Complexity of managing data quality

In the health sector we run into the problem of data integration, since the data we analyze comes from different sources and varies a lot.

1.3.4 Tricky Process or converting big data into valuable insight

It is very difficult to turn this huge amount of big data into expected information

1.3.5 Privacy concern

In the healthcare sector people wouldn't like to share their health condition and diagnosis with others.

1.3.6 Can be a support system but not full automated system

Machine learning models can not be fully automated in replacement with real physicians. Because the decisions are sensitive and lives depend on it.

1.4 Contributions

This thesis provides several insights regarding machine learning techniques on healthcare systems about different disease prediction. We highlight our main contributions here:

1. **Utilizing different classification models**: We used 7 known machine learning classifiers (Logical Regression, Support Vector Machine (SVM) model, Naive Bayes Model, K- Nearest Neighbour, Decision Tree, XGBoost, Neural Network model). Then according to their accuracy we used an accuracy weighted ensemble method which performs better than any individual classifier.

2. **Making combined chronic disease prediction Model**: We developed a central model. Here we sent the individual pre processed numerical datasets of heart, lungs, kidney, liver diseases individually and got individual predictions for each disease. So, it acts as a combined model for these predictions. And as ensemble methods work better, we can use these for other disease predictions as well.

3. **Using COVID - 19 patients data to predict the risk hypothesis of chronic disease**: In our model we passed 281 COVID patients data to predict different chronic disease probability based on different haematological parameters after diagnosed with COVID - 19 and after recovered. Our model can be used now to clinically make predictions and see the actual outcomes which will help to further research in predicting risk of chronic disease of COVID - 19 recovered patients.

4. **Resource Utilization based on fatality prediction**: Our Model can also predict the severity of COVID - 19 patients based on their current condition. And it would help to utilize the remaining resources[13] properly for better healthcare management.

2 Background study / Literature Review

2.1 COVID's impact on chronic diseases

Three different coronaviruses of the same types are Middle East Respiratory Syndrome (MERS), Severe Acute Respiratory Syndrome (SARS) and Coronavirus Disease 2019 (COVID - 19) [3]. COVID - 19 and two other viruses share some common traits in their presentation, as well as their tendency to progression and serious illnesses characterized by high levels of illness and death.

However, comparisons of the three viral diseases also reveal some differences in their clinical manifestations and also in complications, which suggests a variability in the prediction of disease procedure. The narration of the review clearly describes the kidney lung, gastrointestinal, hematologic, heart, hepatic and neurological issues which are associated with these three respiratory diseases or syndromes. It also helps to describe the mechanisms of the immune hyperactivation; which particularly cytokine release syndrome; involved in multiple organ damage detected in severe cases of SARS, MERS and also COVID-19.

Some of the complications of these viruses are:

1. **Pulmonary complications**: Pneumonia is a very common disease of these three viral infections. It usually follows the footprint of the first outbreak of flu-like illnesses including malaise, myalgia, flu and cough occurring in SARS, MERS and COVID - 19. In COVID - 19 pandemic, pneumonia has appeared to predict great effects; 25.9% of the total patients used in the dataset were admitted to the intensive care unit (ICU) in hospital.

2. **Cardiac complication**: Heart problems are commonly reported, at least among the MERS affected patients and COVID - 19 affected patients. In the Saudi 15.7% patients developed arrhythmia, and 14.3% developed rhabdomyolysis from the group of 70 MERS patients. Among the patients of COVID - 19, heart problems also appear to be more prevalent.

3. **Renal complications**: Acute kidney injury (AKI) is one of the most frequent in the context of these diseases, especially MERS which causes kidney damage. 42.9% developed AKI during their illness in the Saudi Arabia. 58% of the MERS affected patients are likely to suffer from renal complications. In COVID-19, 2.5% of the total patients are suffering from kidney damage.

4. **Hepatic complications**: In MERS, severe hepatic injuries were reported. 31.4% of the patients in Saudi Arabia, suffered from severe liver failure during their time of illness. The vast majority of these patients showed elevated aminotransferases when they were stayed in ICU. Prevalence has been significantly reduced in COVID - 19 affected patients although reports of liver dysfunction have been reported

5. **Hematologic complication**: Hypercoagulation and Sepsis are very alarming diseases in respiratory-infection patients and so for COVID - 19 patients, these problems are particularly relevant. In the setting of COVID - 19 patients, the current data of patients do not come in handy in the use of prophylactic anticoagulation. This interplay between the inflammatory and thrombotic processes are also important.

2.2 Haematological parameters

A study of haematological parameters and inflammatory indexes which is used as the parameters to distinguish between the COVID - 19 affected patients from the healthy and not affected people and it also helps to predict the severity of COVID - 19 [6].

2.2.1 Core components

The haematological parameters discussed in this paper can be generalized into a system composed of 5 categories. The categories are as follows:

- Lower lymphocytes, platelet-to-lymphocyte ratio (PLR), platelets, eosinophils, neutrophil-to-lymphocyte ratio (NLR), higher delta neutrophil index (DNI),

basophils were found both in COVID-19 affected patients and influenza groups after comparing with the healthy controls.

- Lymphocytes, eosinophils, PLR have made the biggest contribution to help us differentiate between the COVID-19 affected patients from healthy controls of people.
- Higher neutrophils, DNI, leucocytes, lower lymphocytes, PLR and, red blood cells, haematocrit, haemoglobin levels have been found in the severe patients of COVID-19 at the end of the treatment procedure.
- lymphocytes, eosinophils and platelets showed an increasing curve for the nonsevere patients and neutrophils, DNI, NLR and PLR showed a downward trend.
- Eosinophils, platelets and PLR showed an increasing trend for severe patients.

To come to a conclusion, we can use PLR and NLR hematological parameters which will help to differentiate between the COVID - 19 affected patients from healthy people safe from COVID - 19.

2.3 A brief comparison of machine learning techniques of disease predictions

Motivation of this research is that high accuracy is the supreme necessity in the medical sector for disease prediction and diagnosis and reduces the margin of diagnostic errors [1].

Clinical Decision Support System uses classification Techniques, clustering Techniques and ensemble Techniques.

As each individual classifier or the single classifiers has some kind of limitations and there are some trade-offs. For example those trade offs are training time,

accuracy, scalability, robustness etc.

Some machine learning techniques are used in Heart Disease, Breast Cancer, Diabetes, Liver Disease, Hepatitis patients databases. The problem with these approaches are most of them are single classifiers. Every method has some drawbacks and none of the methods are globally superior to others. No single framework can diagnose multiple diseases with high accuracy.

So ensembling is used to maximize accuracy and diversity. Some combination of classifiers are used such as [1]

1. Majority voting
2. Bagging
3. AdaBoost
4. Stacking etc

Different proposed ensemble models result in high disease diagnosis accuracy and each of them gives better accuracy than the last one. Some ensemble approaches are [1]

1. MV5
2. AccWeight
3. FmWeight
4. BagMoov
5. HM-BagMoov

HM-BagMoov gives the best accuracy.

3 Experimentation

3.1 Overview of the work

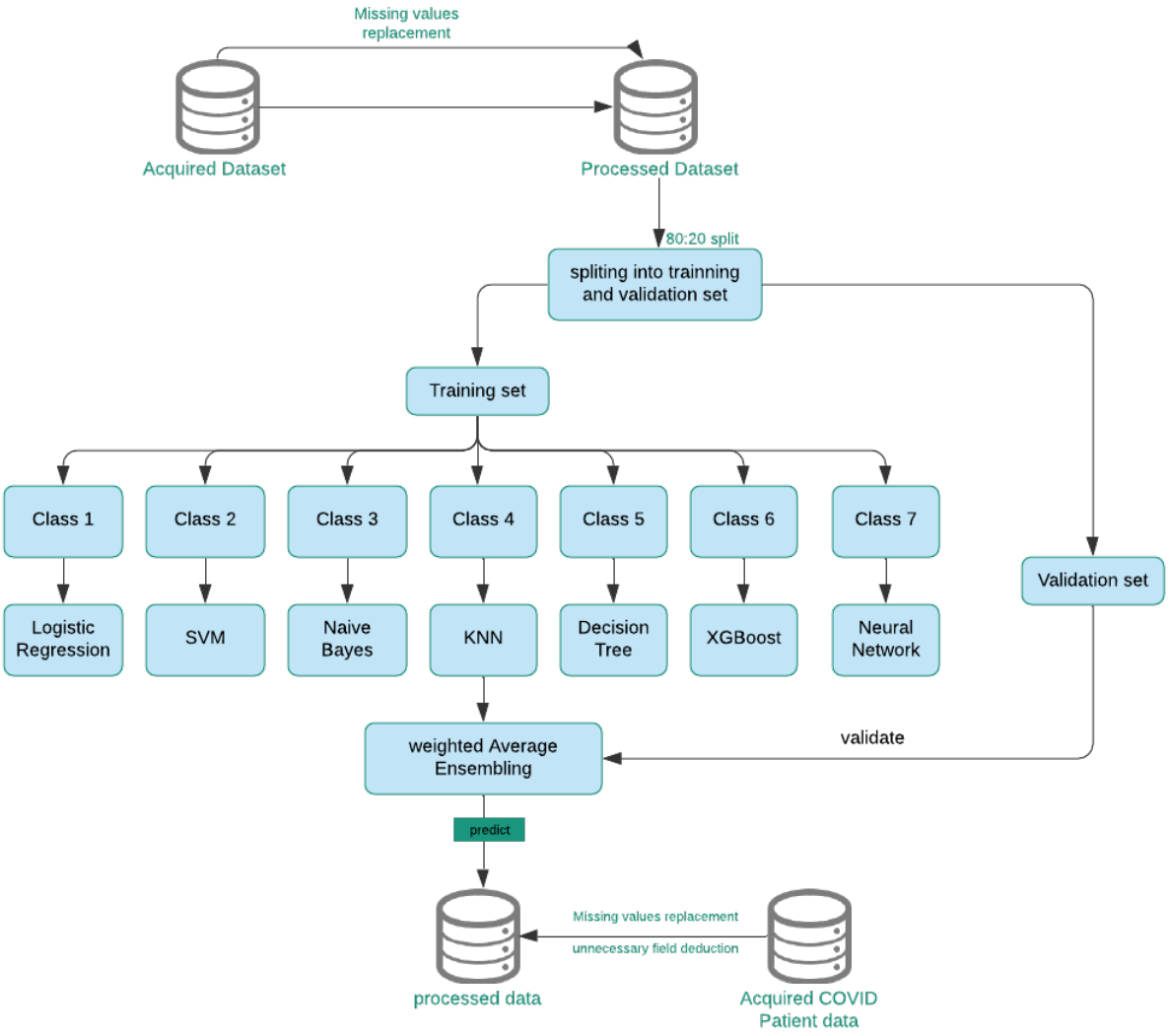

Figure 1: Model Overview

We build a model for different chronic disease predictions. Then passed 281 covid 19 patients data through the model.

3.2 Dataset

For training and testing on the dataset, We needed different types of dataset. We wanted datasets of patients with their medical history and For different disease predictions we used different datasets. Using different dataset, we wanted to apply our model for predicting different specific chronic diseases. .Including heart, liver, COPD, lungs etc. The dataset we have used to train the model goes here.

1. **Cleveland Heart disease data**: [14] The dataset consists of 303 individuals. There are 14 columns in the dataset like age, sex, chest pain type, resting blood pressure, serum cholestrol, fasting blood suger, rasting ecg, max heartrate achieved, excercise induced angina, ST depression induced by excercise relative to rest, Peak exercise ST segment, Number of major vessels (0–3) colored by flourosopy, Thal, Diagnosis of heart disease etc.

2. **Statlog (Heart) Data Set**: [17] This dataset is a heart disease database with 13 attributes like age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar ¿ 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal (where 3 = normal; 6 = fixed defect; 7 = reversable defect).

3. **SPECT Heart Data Set**: [16] The dataset describes diagnosing of cardiac Single Photon Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature patterns were created for each patient. The pattern was further processed to obtain 22 binary feature patterns. The CLIP3 algorithm was used to generate classification rules from these patterns. The CLIP3 algorithm

generated rules that were 84.0% accurate (as compared with cardiologists' diagnoses).

Cleveland, Statlog, SPECT Heart Data Set have similar types of parameters.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1

Figure 2: Heart disease dataset

4. **Indian Medical Liver disease dataset**: [15] This data set contains 416 liver patient records and 167 non liver patient records.The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records.Any patient whose age exceeded 89 is listed as being of age "90".

   Attribute were age of the patient, gender of the patient, TB Total Bilirubin, DB Direct Bilirubin, Alkphos Alkaline Phosphatase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, TP Total Proteins, ALB Albumin, A/G Ratio Albumin and Globulin Ratio, Selector field used to split the data into two sets (labeled by the experts).

5. **BuPa Liver Disease dataset**: [13] The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the dataset constitutes the record of a single male individual.

   Seven attribute here are: mcv mean corpuscular volume, alkphos alkaline phosphotase, sgpt alanine aminotransferase, sgot aspartate aminotransferase, gammagt gamma-glutamyl transpeptidase, drinks number of half-pint

equivalents of alcoholic beverages drunk per day, selector field created by the BUPA researchers to split the data into train/test sets.

6. **COVID Patients Medical Records**: [6] This is a dataset gathered by Cambridge University Press. Here we have data of 281 patients. And different types of features set which 74 in numbers. Half of the features are extracted from patients while getting admitted and half of them are extracted after the treatment has been done. All the features there are listed below:

Table 1: Features table

1	White blood cells	The cells of the immune system that are involved in protecting the body against both infectious disease and foreign invaders
2	Red Blood Cell	A type of immune cell that is one of the first cell types to travel to the site of an infection
3	Neutrophils	A small white blood cell (leukocyte) that defends the body against disease
4	Lymphocyte	A small white blood cell (leukocyte) that defends the body against disease
5	Monocytes	A white blood cell that has a single nucleus and can take foreign material
6	Eosinophils	A type of disease-fighting white blood cell
7	Basophils	White blood cells from the bone marrow that play a role in keeping the immune system functioning correctly
8	Hemoglobin	Oxygen-carrying component of red blood cells
9	Hematocrit	The proportion of the blood that consists of packed red blood cells
10	MCV	Blood test measures the average size of your red blood cells, also known as erythrocytes

11	MCHC	Measure of the average concentration of hemoglobin inside a single red blood cell
12	MCH	The average quantity of hemoglobin present in a single red blood cell
13	RDW	A measurement of the range in the volume and size of your red blood cells (erythrocytes)
14	Platelets	help form blood clots to slow or stop bleeding and to help wounds heal
15	MPV	small blood cells that are essential for blood clotting, the process that helps you stop bleeding after an injury
16	PCT	used in a variety of clinical settings including primary care, emergency department and intensive care
17	PDW	Used as regular parameter in blood routine examination
18	LUC	differential count parameter measured by certain routine hematology analyzers and reflects activated lymphocytes and peroxidase-negative cells
19	NRBC	Reflects high production of erythropoietin; means erythropoietin stimulates fetal hematopoietic system, mainly in bone marrow
20	DNI	Immature granulocyte fraction provided by a blood cell analyser
21	NLR	Calculated as a simple ratio between the neutrophil and lymphocyte counts measured in peripheral blood
22	PLR	A test that predicts whether cardiac output will increase with volume expansion
23	GFR	A blood test that checks how well your kidneys are working
24	ALT	A test is typically used to detect liver injury

25	AST	An enzyme found in cells throughout the body but mostly in the heart and liver and, to a lesser extent, in the kidneys and muscles
26	LDH	An enzyme involved in energy production that is found in almost all of the body's cells, with the highest levels found in the cells of the heart, liver, muscles, kidneys, lungs, and in blood cells
27	CK	An enzyme found in the heart, brain, skeletal muscle, and other tissues
28	PT	A blood test that measures the time it takes for the liquid portion (plasma) of your blood to clot
29	aPTT	A screening test that helps evaluate a person's ability to appropriately form blood clots
30	INR	A calculation based on results of a PT and is used to monitor individuals who are being treated with the blood-thinning medication
31	D-dimer	One of the test to see the Coagulation Profile
32	CRP	Checks for inflammation in the body which can be caused by infection, injury, or chronic disease
33	IL-6	An endogenous chemical which is active in inflammation, and in B cell maturation
34	Ferritin	A globular protein complex consisting of 24 protein subunits forming a nanocage

Patient_number	(Age)	(Gender)	(groups)	(Diabetes)	(hypertension)	(coronary artery disease)	(COPD)	COVID-19 Severity	admission WBC	admission Neutrophils	admission Lymphocytes	admission monocytes	admission eosinophils
C1	83	1	2	0	1	1	0	1	5630	3670	1280	450	10
C2	75	0	2	1	1	1	1	0	5750	3610	1310	500	180
C3	57	1	2	1	0	0	0	0	6170	3010	2740	240	50
C4	57	0	2	0	1	0	1	1	3310	1740	1150	190	20

Figure 3: covid-19 patients data

admission RBC	admission hemoglobin	admission hematocrit	admission MCV	admission MCH	admission MCHC	admission RDW	admission platele	MPV	PCT	PDW	admission LUC	admissin LUC%	admission NRBC	admission NRBC%	admission DNI
5.06	15.5	44.8	88.5	30.6	34.6	13.4	275	8.2	0.22	45.5	0.09	1.4	0	0	0.1
6.19	16.7	49.6	80.2	27	33.7	12.9	279	8.3	0.23	63.5	0.2	2.5	0	0	0.1
5.23	15.5	44.6	85.2	29.7	34.8	13.1	180	9.1	0.16	43.1	0.09	1.5	0	0	0.1

Figure 4: covid-19 patients data

admission NLR	admission	admission	admission ALT	LDH	CK	PT	aPTT	INR	D-dimer	CRP	IL-6	Ferritin	end of treatment W	Neutroph	Lymphoc	Monocyte	Eosinophi	Basophils	RBC	Hemoglob
2.86	0.17	71	23	35	273	163	12.9	27.2	1.1	1.49	36	330	5020	3360	1070	340	40	20	4.55	14.3
2.75	0.19	72	19	17	258	77	12.3	20.9	1.05	0.29	22	58.3	5780	3510	1640	280	190	30	4.11	10.8
1.09	0.08	91	25	11	190	142	12.3	24.6	1.05	0.19	0.7	16	7140	3450	3060	310	130	90	5.19	14.8
2.00	0.17	98	112	74	266	30	11.3	21.1	1	0.45	0.6	407	5650	3260	1610	420	200	20	4.16	12.4

Figure 5: covid-19 patients data

3.3 Data Preprocessing

Here we are dealing with medical history data. There are always a few issues with the medical history dataset. Because these datas are not something that can be gathered automatically. Rather these are gathered or generated in many random processes. Some are collected from medical diagnosis reports, some are collected from hand made entries, few are with some automated process.

The problem occurs with this kind of medical history datas are:

1. Medical data is often **heterogeneous** by nature (different types of features.
2. Very susceptible to **data imbalance** (the class of interest is usually under-represented)
3. To **missing data** (data is generated nearly every second, handled by several different people within the institutions and saved in different formats. . .).

To eradicate all these issues we have come up with the below solutions.

1. **Missing values replacement**: Missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Hematocr	MCV	MCHC	MCH	RDW	platelets	MPV	PCT	PDW	LUC	LUC%	NRBC	NRBC%	DNI	NLR	PLR
41.5	91.2	31.5	34.6	13.1	266	7.1	0.19	53.1	0.19	3.8	0	0	0.1	3.14	0.25
31.8	77.5	26.2	33.8	14.5	314	7.6	0.24	56.8	0.12	2.1	0	0	0.1	2.14	0.15
43	83	28.5	34.3	13.7	213	6.6	0.14	45.4	0.1	1.4	0	0	0.1	1.13	0.07
36.2	87	29.8	34.2	12.5	334	7.6	0.25	62	0.13	2.3	0	0	0.1	2.02	0.21

Figure 6: covid-19 patients data

In several fields there were no values. And those were creating null values in the processing and predication. We had to take the mean value of that specific column and replace those null values with that mean value.

2. **Outlier Detection and removal**: Few values are there such that they are quite deviating from the mean value of the column. Those few specific values are the outlier and they are affecting the accuracy of the model. Those values have been deducted.

3. **Normalizing the dataset**: Datas are varying in a large range. We normalize them and bring them within a small region of o to 1. It helps the model to predict and learn well.

   Normalized data,

$$x' = (x - x_{min}) \,/\, (x_{max} - x_{min}) \tag{1}$$

3.4 Feature Analysis

We have selected important parameters from different research papers and with help of doctors for different chronic disease predictions. For heart disease prediction, we found that CBC (Specially fibrinogen, c-reactive protein [19], ECG, ECO can help to identify the diseases. We also found after COVID-19 there are significant changes in these parameters [23]. And there are significant heart disease prediction models based on this parameters [20] [8] [21] [2] For Lungs disease we found that pneumonia [4] [2] is common in COVID-19 and often it results in COPD. X-ray [5], city-scan [7] images play an important role to identify the lung diseases.

3.5 Classification Algorithms

In this paper, we took help of a number of classification models which eventually supported our ensembled model. All these known classification models have different approaches to make.

1. **Logistic regression**: Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between the chronic diseases taking place or not and one or all the clinical features we could have generated.

2. **Support Vector Machine (SVM) model**: A support vector machine (SVM) is a machine learning model which works on labeled datasets only. And especially when we need algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for diseases, they're able to categorize the output into the result of the diseases being attacked.

3. **Naive Bayes Model**: It is a classification technique based on Bayes' Theorem with an assumption of independence among the diseases. Here, the Naive Bayes classifier simply assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

4. **K- Nearest Neighbour**: K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It takes the nearby possible outcomes and brings out the outcome prediction by seeing the output of the close neighbour.

5. **Decision Tree**: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

So, we make decisions by breaking all our dataset into smaller pieces while at the same time it is associated with a decision tree. And as the final result we get to some positive or negative answers. In this case, it would be affected by diseases and not affected by diseases. Decision tree can handle both numeric and categorical data and we have it both.

The core algorithm used here is ID3. This is a top down greedy search approach. The increasing and decreasing of entropy plays a vital role in this calculation.

6. **XGBoost**: XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

7. **Neural Network model**: Neural Networks are methods for predicting things. They provide a simple non-linear relationship between the variable and predicator. They are some layers in a neural network model. Some are input output layers. And others hidden layers. These hidden layers can have any nodes.

3.6 Model Structure

The model is based on a weighted average ensembling method. We take the impacts of different algorithms and put some weight on them according to their performance. And then we found out real predictions, contributed by each of the basic classifiers.

Then we used the accuracy weighted ensemble to get a better prediction model. All these predict in a different way, based on different features. After getting the predicted outputs from those models they go through the ensembling learning model.

1. **Ensembling method**: Ensembling models are based on a very straightforward concept. They take multiple models and then integrate all the model

performance to improve the output.

There is a problem with deep learning methods. That they provide increased flexibility and this is a problem because this makes them sensitive to the features of training data and might find different weights each time they are trained. And thus different predictions come up.

To solve this problem, ensembling methods are being used. Which enables multiple models instead of a single model. This not only avoids the variation in prediction but also accurates the prediction.

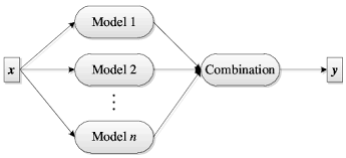

Figure 7: Ensembling Model

Ensembling methods can be based on two different methods:

(a) Voting

(b) Average

Both of these methods are the easiest ensembling methods. Voting is used for classification and average is used for regression analysis.

In any method, we need to create multiple classification or regression models using some dataset, which will train them. Then we need to split the full training dataset into different splits and apply the same or different algorithm.

But while merging the outputs, the dilemma comes which splitted train data gonna play the anchor role while dealing with the final prediction. Here we come up with other concepts.

2. **Majority Voting**: Majority voting is one way to find out which model to give priority and which to not. Here each of the models makes predictions on every test set. And finally which of the model gets more than half of

the vote gets relied on that model. If none of the models gets more than half votes, me might say none stable prediction is being made and we should drop the idea of ensembling for that instance. But if you really need to use that you need to get the highest vote achiever.

3. **Weighted Voting**: In weighted voting, none gets to play a solo anchor. Rather all of the model gets to contribute here. But we just increase the importance of one model here. Here we try founding out the count of the predictions of the better models multiple times. FInding a reasonable set of weights can be tricky. But we are allowed to do it as it comes up.

   We might come up with gradient descent techniques for getting with the weights.

4. **Simple Averaging**: In this method, we do the averaging for every instance, This method often reduces overfit and creates a smoother regression model. The following pseudocode code shows this simple averaging method

```
final_predictions = []
for row_number in len(predictions):
    final_predictions.append(
        mean(prediction[row_number, ])
    )
```

Figure 8: Algorithm

This learning model carries weights for different classes or models. Giving different measures of priorities to different classes. This includes weighted voting and this carries the contribution of the classes in our central model.

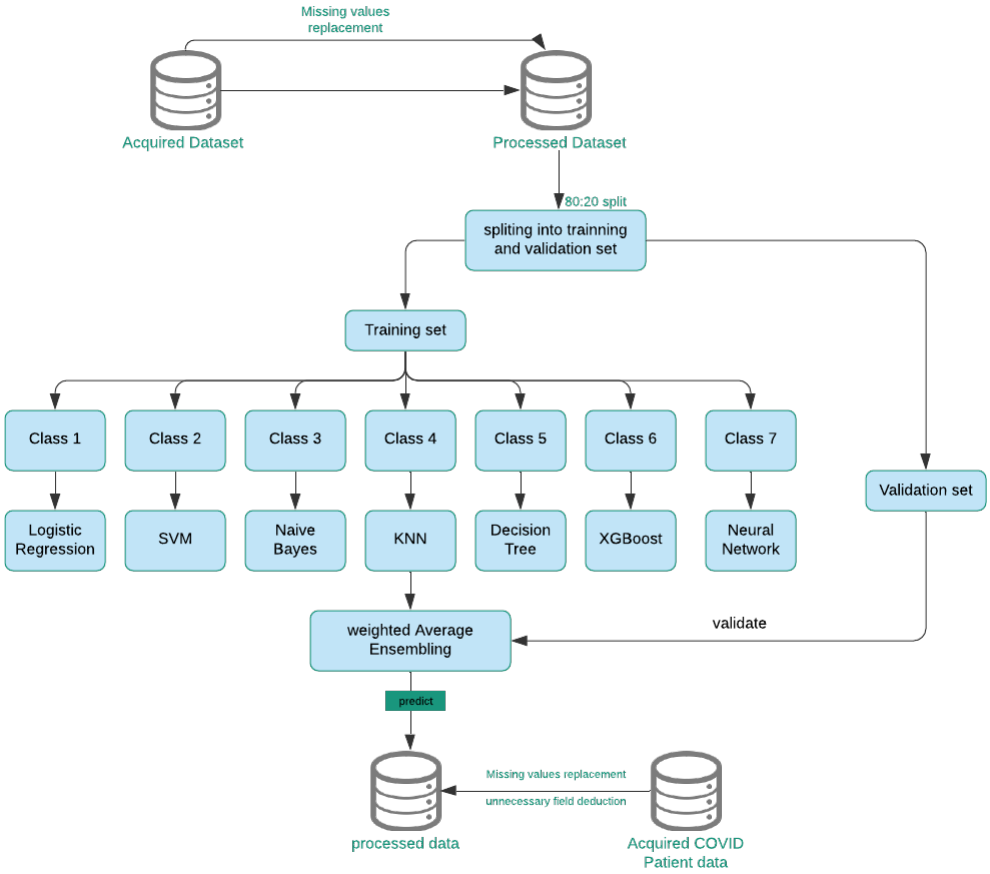

Figure 9: Model overview

3.7 Final Model

Then using our ensemble method we built heart, liver, kidney disease prediction model. Then we passed 281 covid-19 patients data to this model to check the probability of developing chronic disease.

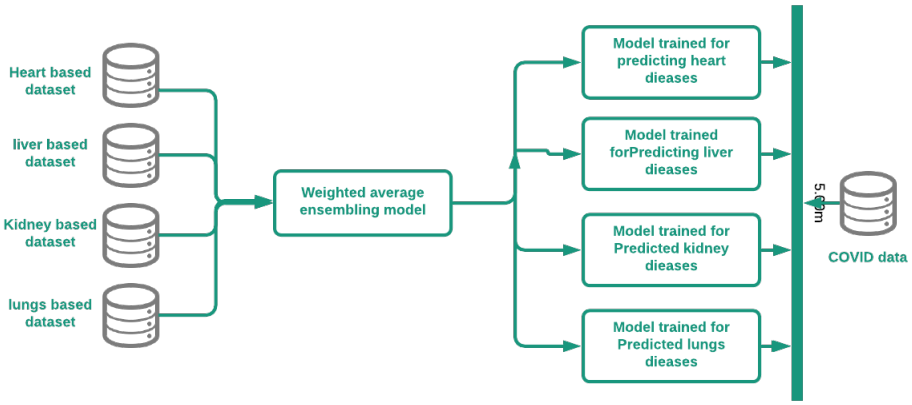

Figure 10: Using Final Model

4 Results Analysis

After training our developed weighted average model, we applied those on different dataset for finding. We have developed an accuracy weighted average model from suitable classification classifiers. In our outcome, we have come up with different diseases.

1. **Heart Diseases prediction**: On a heart disease data, where we applied our model on the Cleveland Heart disease data. We have got an output of 89.92% from the ensembled model. All the other models here, performed accordingly.

Table 2: Comparison table for Heart

	Logical Regression	SVM	Naive Bayes	Decision tree	XGBoost	Neural Network	KNN
Accuracy	80.67%	80.25%	85.29%	88.24%	68.82%	77.31%	68.07%
F1 Score	0.80%	0.80%	0.85%	0.88%	0.88%	0.77%	0.67%

After the results, we see that the developed algorithm out performed the basic algorithms.

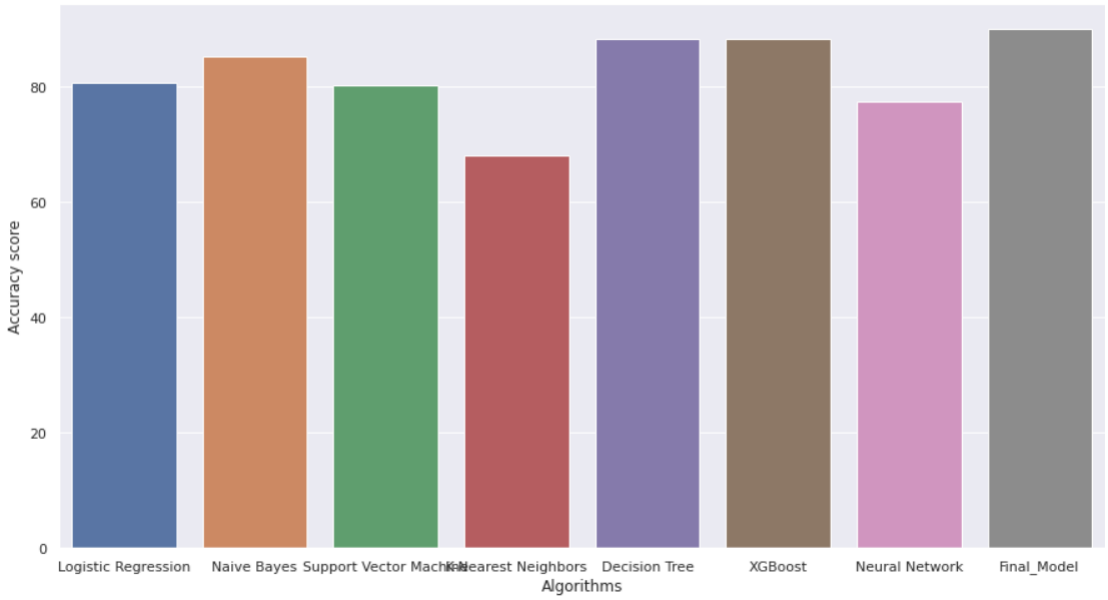

Figure 11: Analysis of Heart Diseases prediction

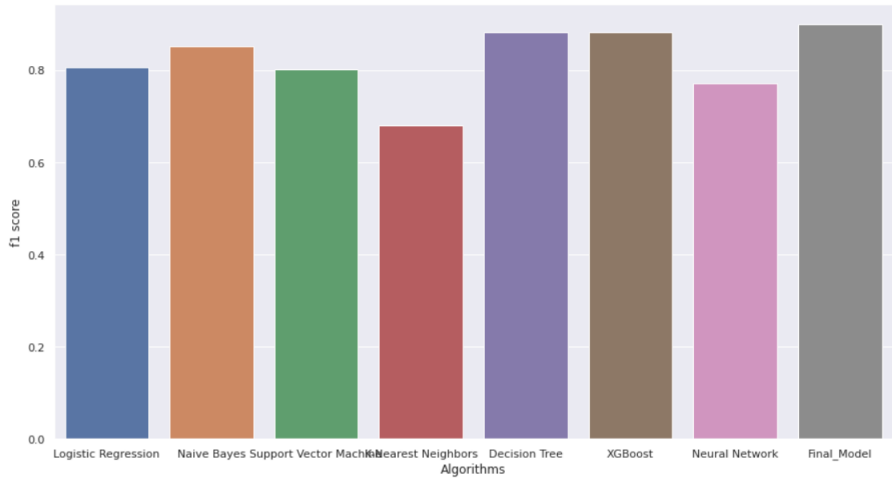

Figure 12: F1 score of Heart Diseases prediction

2. **Liver Disease Prediction**: For liver disease prediction, we applied our dataset predictions Our liver disease prediction model accuracy is 74% (with 584 patients data from Indian liver patients dataset) 73%(on BUPA liver patients dataset).

Table 3: Comparison table for Liver

	Logical Regression	SVM	Naive Bayes	Decision tree	XGBoost	Neural Network	KNN
Accuracy	63.98%	56.99%	53.23%	70.43%	68.82%	56.99%	68.82%
F1 Score	0.56%	0.35%	0.48%	0.70%	0.67%	0.72%	0.68%

So far in the published paper accuracy is 73%(only using Neural network), and 72.5%(only using SVM model).

Though our main goal is to predict COVID patients chronic disease probability, we had to build our own model but it supported the accuracy).

Then we passed 282 COVID patients data to our model and from that we found that they have 32% possibility of gaining liver disease.

The model accuracy comparison with F1 score is given below as well:

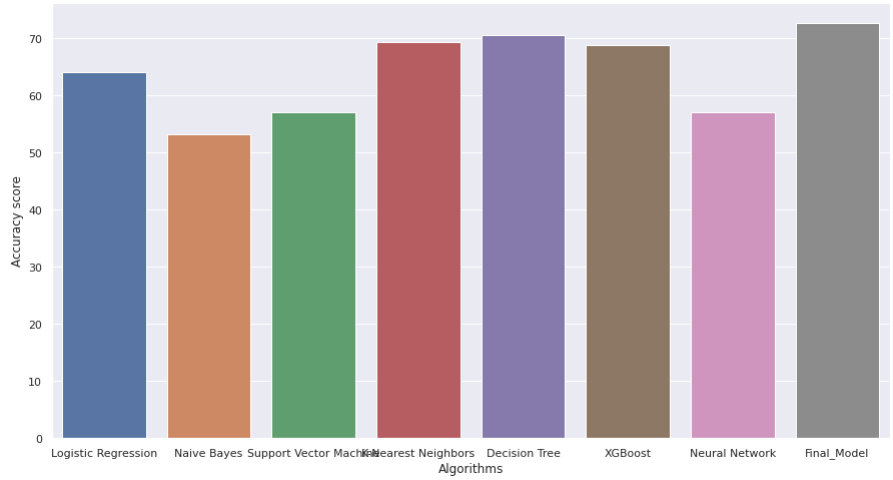

Figure 13: Accuracy score of liver disease prediction

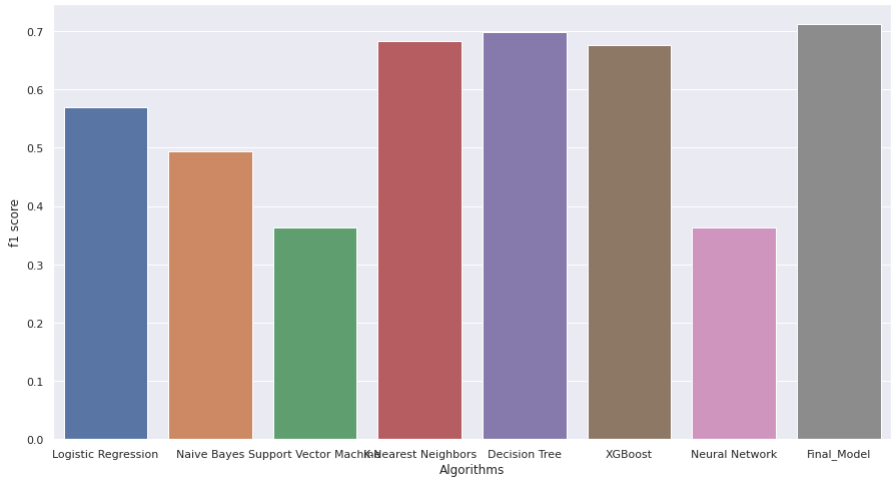

Figure 14: F1 score of liver disease prediction

Kidney Disease Prediction: On a kidney disease data, where we applied our model on the kidney disease data. We have got an output of 96.88% from the ensembled model. All the other models here, performed accordingly

Table 4: Comparison table for Kidney

	Logical Regression	SVM	Naive Bayes	Decision tree	XGBoost	Neural Network	KNN
Accuracy	96.88%	96.88%	99.6%	96.88%	68.82%	59.38%	87.60%
F1 Score	0.805%	0.80%	0.85%	0.88%	0.8816%	0.77%	0.68%

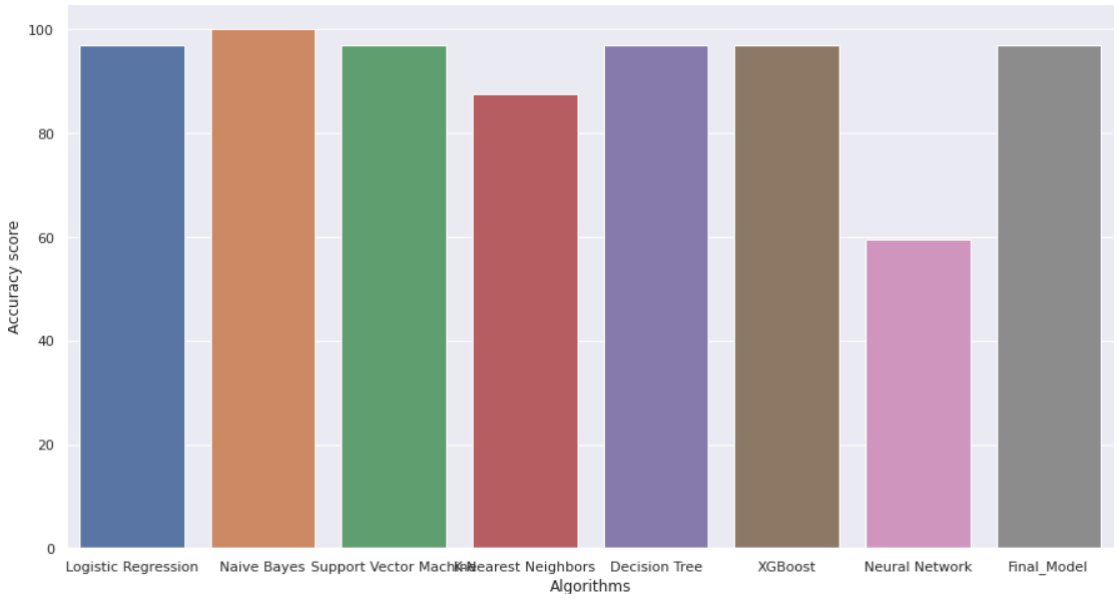

Figure 15: Accuracy score for kidney disease prediction

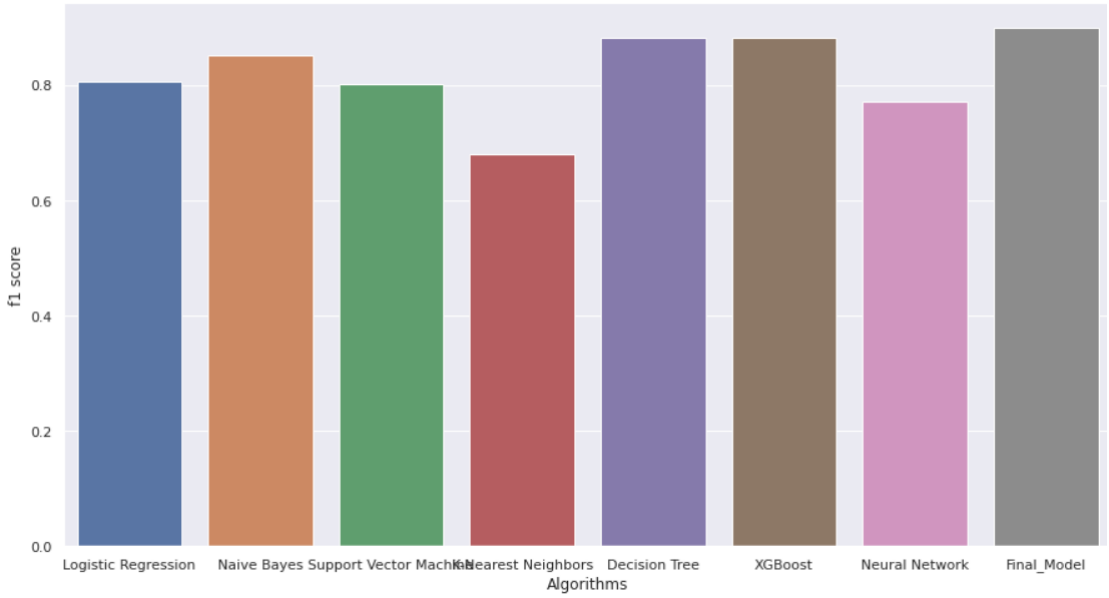

Figure 16: F1 score for kidney disease prediction

5 Discussions and Conclusions

In brief, we developed a central model using some popular classification algorithms and combining them into accuracy weighted average ensemble technique. Then we built the predictive model for some chronic diseases (heart, kidney, liver, lungs disease). We found that our central model works quite well in different individual disease predictions. Then we passed 281 COVID - 19 patients data and now our models are ready for predicting chronic disease of COVID - 19 patients. Now it can be used in different researches to predict the probability of chronic diseases after recovery from infectious disease (i.e. COVID - 19) clinically. And the outcome of clinical results will help the patients to take predictive measures. Thus casualties can be minimized.

6 Challenges and Future Works

1. **Collaborating with Hospitals**: As we are working with medical data which are both sensitive and have privacy issues. So, there are a lot of

challenges. But millions of COVID - 19 patients admitted and recovered from hospitals, so, working with their data would be challenging as well as a great opportunity for taking this research further.

2. **Developing the central model for image data as well**: Our model so far works well with numeric data. But ECG, ECHO, CT-Scan, X-ray works are also essential for these chronic disease predictions and also identifying the severity. We are working on developing an all in one model which works both for image numeric data as well and acts as an assisting model in giving treatment.

3. **Working with other infectious disease**: Our work was so far based on chronic disease prediction for COVID - 19 recovered patients. There are similar infectious disease SARS, MARS also different virus affected diseases. Now we will be able to work on predicting long term effects of different infectious disease as well as the increased or decreased probability of chronic diseases after infectious disease recovery.

References

[1] Saba Bashir, Usman Qamar, and Farhan Hassan Khan. Intellihealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of biomedical informatics*, 59:185–200, 2016.

[2] Zhichao Feng, Qizhi Yu, Shanhu Yao, Lei Luo, Wenming Zhou, Xiaowen Mao, Jennifer Li, Junhong Duan, Zhimin Yan, Min Yang, et al. Early prediction of disease progression in covid-19 pneumonia patients with chest ct and clinical characteristics. *Nature communications*, 11(1):1–9, 2020.

[3] Jad Gerges Harb, Hussein A Noureldine, Georges Chedid, Mariam Nour Eldine, Dany Abou Abdallah, Nancy Falco Chedid, and Wared Nour-Eldine. Sars, mers and covid-19: clinical manifestations and organ-system complications: a mini review. *Pathogens and Disease*, 78(4):ftaa033, 2020.

[4] Elisa Grifoni, Alice Valoriani, Francesco Cei, Vieri Vannucchi, Federico Moroni, Lorenzo Pelagatti, Roberto Tarquini, Giancarlo Landini, and Luca Masotti. The call score for predicting outcomes in patients with covid-19. *Clinical Infectious Diseases*, 72(1):182–183, 2021.

[5] Yan Han, Chongyan Chen, Ahmed H Tewfik, Ying Ding, and Yifan Peng. Pneumonia detection on chest x-ray using radiomic features and contrastive learning. *arXiv preprint arXiv:2101.04269*, 2021.

[6] Sumeyye Kazancioglu, Aliye Bastug, Bahadir Orkun Ozbay, Nizamettin Kemirtlek, and Hurrem Bodur. The role of haematological parameters in patients with covid-19 and influenza virus infection. *Epidemiology & Infection*, 148, 2020.

[7] Wassim W Labaki, Carlos H Martinez, Fernando J Martinez, Craig J Galbán, Brian D Ross, George R Washko, R Graham Barr, Elizabeth A Regan, Har-

vey O Coxson, Eric A Hoffman, et al. The role of chest computed tomography in the evaluation and management of the patient with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 196(11):1372–1379, 2017.

[8] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019.

[9] S Muthuselvan, S Rajapraksh, K Somasundaram, and K Karthik. Classification of liver patient dataset using machine learning algorithms. *International Journal of Engineering & Technology*, 7(3.34):323–326, 2018.

[10] Helena Nyblom, Ulf Berggren, Jan Balldin, and Rolf Olsson. High ast/alt ratio may indicate advanced alcoholic liver disease rather than heavy drinking. *Alcohol and alcoholism*, 39(4):336–339, 2004.

[11] World Health Organization. Coronavirus disease (covid-19) pandemic. *https://www.who.int/emergencies/diseases/novel-coronavirus-2019*, 2021.

[12] Nenad Petrović. Simulation environment for optimal resource planning during covid-19 crisis. In *2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, pages 23–26. IEEE, 2020.

[13] UCI Machine Learning Repository. Bupa liver disease dataset. *https://archive.ics.uci.edu/ml/datasets/liver+disorders*.

[14] UCI Machine Learning Repository. Cleveland heart disease data set. *https://archive.ics.uci.edu/ml/datasets/heart+disease*.

[15] UCI Machine Learning Repository. Indian liver patient dataset. *https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)*.

[16] UCI Machine Learning Repository. Spect heart data set. *https://archive.ics.uci.edu/ml/datasets/spect+heart*.

[17] UCI Machine Learning Repository. Statlog (heart) data set. *http://archive.ics.uci.edu/ml/datasets/statlog+(heart)*.

[18] Amy G Shah, Alison Lydecker, Karen Murray, Brent N Tetri, Melissa J Contos, Arun J Sanyal, Nash Clinical Research Network, et al. Comparison of noninvasive markers of fibrosis in patients with nonalcoholic fatty liver disease. *Clinical gastroenterology and hepatology*, 7(10):1104–1112, 2009.

[19] Amit Kumar Shrivastava, Harsh Vardhan Singh, Arun Raizada, and Sanjeev Kumar Singh. C-reactive protein, inflammation and coronary heart disease. *The Egyptian Heart Journal*, 67(2):89–97, 2015.

[20] Marjia Sultana, Afrin Haider, and Mohammad Shorif Uddin. Analysis of data mining techniques for heart disease prediction. In *2016 3rd international conference on electrical engineering and information communication technology (ICEEICT)*, pages 1–5. IEEE, 2016.

[21] Abhishek Taneja et al. Heart disease prediction system using data mining techniques. *Oriental Journal of Computer science and technology*, 6(4):457–466, 2013.

[22] Liam Townsend, Adam H Dyer, Karen Jones, Jean Dunne, Aoife Mooney, Fiona Gaffney, Laura O'Connor, Deirdre Leavy, Kate O'Brien, Joanne Dowds, et al. Persistent fatigue following sars-cov-2 infection is common and independent of severity of initial infection. *Plos one*, 15(11):e0240784, 2020.

[23] Clyde W Yancy and Gregg C Fonarow. Coronavirus disease 2019 (covid-19) and the heart—is heart failure the next chapter? *JAMA cardiology*, 5(11):1216–1217, 2020.

[24] Li Zuo, Ying-Chun Ma, Yu-Hong Zhou, Mei Wang, Guo-Bin Xu, and Hai-Yan Wang. Application of gfr-estimating equations in chinese patients with chronic kidney disease. *American journal of kidney diseases*, 45(3):463–472, 2005.