Prediction and Detection of Infectious Disease through Machine Learning

[1]Manoj Kumar Singh, [2] Dr. Kishan Pal Singh, [3] Dr. Devendra Kumar
[1] Research Scholar Mangalayatan University Aligarh India, [2] Professor Mangalayatan University Aligarh India,

[3] Professor ABES Engineering College Ghaziabad India

[1]singhmanojmit@gmail.com, [2] drkishan.mu@gmail.com, [3] devendra.arya@gmail.com

Article History: Received:05.02.2023 Revised:20.03.2023 Accepted: 05.06.2023

Abstract- Infectious diseases are caused by organisms such as bacteria, viruses, fungi or parasites. They live in and on our bodies. Learning about infectious diseases provides a better understanding of human pathogens and the need for treatments and immunizations. These critical tasks may be carried out in a single clinic by infectious disease clinicians (disease physicians) with specialised training or with exceptional precision. Machine learning (ML) is employed in a variety of fields, including education and healthcare. It is utilised extensively in healthcare. In healthcare, ML is utilised to tackle a variety of challenges. Developing a machine learning model, training it on the dataset, and incorporating unique patient information may aid in prediction of the disease. The forecast result will be based on the data provided and so will be unique to that individual. Corona virus is an illness with no specific treatment. There are several therapies available for it, however there are no well-defined therapy processes. The goal of artificial intelligence (AI) is to replicate human cognitive processes. In this study we aimed to detect and predict of infectious disease, perform automated learning to predict and accurately detect infection and construct the model that can perform in real time aspects.

Keyword: COVID-19, Infectious Illness Therapy, Machine Learning, SARS-CoV-2, infectious diseases, ML-Algorithm, Probabilistic Fuzzy Logic, MATLAB, Python.

1 INTRODUCTION

Corona viruses

Infectious diseases (Bloom &Cadarette, 2019) are disorders caused by organisms such as bacteria, viruses, fungi or parasites (Agrebi&Larbi 2020). Many organisms live in and on our bodies. There are a wide variety of creatures that call our body home. They rarely cause harm and sometimes even aid. However, certain microbes may cause illness under specific circumstances. It is possible to spread certain infectious illnesses from one person to another (Morse, 2001). Bugs and other animals may be responsible for spreading certain diseases. There are many different kinds of illnesses, and you can catch some from eating or drinking something that is tainted, or from being exposed to an organism in the wild. Infection symptoms may range from mild to severe, depending on the organism that’s causing the illness. Some infections are mild enough to be treated at home with rest and over-the-counter medication, while others are so severe that only hospitalization will do. Vaccines are an effective way to guard against contracting infectious illnesses like measles and chickenpox (Cohen et al., 2021). Hand-washing, both often and thoroughly, provides an additional line of defense against the spread of illness. The symptoms of an infectious illness vary depending on the ailment. Symptoms that may indicate a variety of different infectious illnesses are:

- Fever
- Diarrhea
- Fatigue
- Muscle aches
- Coughing

. Infectious diseases can be caused by:

- **Bacteria.** These one-cell organisms are responsible for illnesses such as strep throat, urinary tract infections and tuberculosis.
- **Viruses.** Even smaller than bacteria, viruses cause a multitude of diseases ranging from the common cold to AIDS.
- **Fungi.** Many skin diseases, such as ringworm and athlete’s foot, are caused by fungi. Other types of fungi can infect your lungs or nervous system.

- **Parasites.** Malaria is caused by a tiny parasite that is transmitted by a mosquito bite. Other parasites may be transmitted to humans from animal faces.

1.1Corona

COVID-19 is an illness caused by the SARS-CoV-2 corona virus (Ortiz-Prado et al., 2020). The World Health Organization first became aware of this novel virus on December 31, 2019(Sohrabi et al., 2020), after a report of a cluster of cases of ‘viral pneumonia’ in Wuhan, People’s Republic of China. Corona virus disease (COVID-19) is a virus-borne infection caused by the SARS-CoV-2 viral (Tran et al., 2021). Most patients infected with the virus will have mild to severe respiratory sickness and will recover without needing any specific therapy. Some, though, may get very sick and need medical treatment. People over the age of 65, as well as those with underlying medical disorders such as cardiovascular disease, diabetes, chronic respiratory disease, or cancer, are at a higher risk of developing severe illness. Anyone of any age may get very ill or die as a result of COVID-19. Being thoroughly educated on the sickness and how the virus spreads is the greatest method to avoid and slow down transmission. Stay at least one metre away from people, wear a well fitting mask, and wash your hands or use an alcohol-based rub often to protect yourself and others from infection. When it’s your time, get vaccinated and follow local recommendations. When an infected person coughs, sneezes, speaks, sings, or breathes, the virus may spread in microscopic liquid particles from their mouth or nose (Stadnytskyi et al., 2021). These particles vary in size from big respiratory droplets to tiny aerosols. If you feel ill, it is critical to adopt respiratory etiquette, such as coughing into a flexed elbow, and to remain at home and self-isolate until you recover.

Symptoms of the new Corona virus: - The most common symptoms include:

Fever

Cough

Shortness of Breath and Difficulty breathing

1.2 Machine learning and Infectious Disease

Machine learning is employed in a variety of fields, including education and healthcare (UmaMaheswaran et al., 2022). Better computer power and the availability of datasets on open-source repositories have enhanced the application of machine learning as technology has advanced. Machine learning is utilized extensively in healthcare. The healthcare industry generates a vast quantity of data in the form of photographs, patient data, and so on, which helps in identifying trends and making predictions. In healthcare, machine learning is utilized to tackle a variety of challenges. Heart illness is determined by the individual, and the severity of heart disease varies from person to person. As a result, developing a machine learning model, training it on the dataset, and incorporating unique patient information may aid in prediction. The forecast result will be based on the data provided and so will be unique to that individual. Type 2 diabetes is a condition that may be avoided by maintaining a healthy weight and lifestyle. Corona virus is an illness with no specific treatment (Sankar et al., 2020). The corona virus 2019 (COVID-19) emerged in China (Shi et al., 2020). There are several therapies available for it, however there are no well-defined therapy processes. The goal of artificial intelligence (AI) is to replicate human cognitive processes. It is ushering in a paradigm change in healthcare, fueled by the growing availability of healthcare data and the fast advancement of analytics tools. Many models for automated diagnosis of illnesses such as cancer, COVID-19, and diabetes have recently been created. Many academics have recently begun to use machine learning models for real-time illness detection by constructing Smartphone applications. Even some mobile applications have been created that may forecast the likelihood of a certain illness and offer a diagnosis depending on the individual’s health parameters. However, effective early-stage diagnosis remains an unsolved challenge. Many academics have recently begun to use deep-learning models to achieve much higher performance than machine learning models.

2 Research Background

Wu et al. (2020), Based on clinically accessible blood test results, machine learning-based quick and reliable identification of COVID-19 infection. This discovery has the potential to capture the fundamental patterns of these regular parameters, enabling routine blood tests to influence the value of warning illnesses. Following several evaluations, it is of exceptional benefit to be used for large-scale initial screening on suspected COVID-19 patients in order to reduce the virus’s risk, particularly for confirmed patients who are promptly isolated and treated, which is currently known to be the most effective response. Although additional clinical testing of the suggested instrument is required, the approach provides some new insights into the early detection of COIVD-19 infection and how to cope with disaster situations owing to the terrible features of human-to-human transmission.

Bogu et al. (2021), Deep learning is being used to identify COVID-19 using wearable data. By 2020, COVID-19 is anticipated to kill up to 17 million people worldwide. In this experiment, they constructed a deep-learning system that utilised wearable data such as abnormal resting heart rate to predict COVID-19 before symptoms emerged. Time-series datasets from 25 COVID-19+ patients, 11 non-COVID-19 patients, and 70 healthy persons were used to train the deep-learning system. To the best of our knowledge, this is the first deep-learning model to identify an early viral infection using wearable data at the user level. In order to quantify performance indicators, this study expands on our previous phase-1 research by taking into consideration the data’s unexpected behaviour and time-series structure, as well as its restricted size and lack of data labels. As the number of COVID-19 patients grows, a real-time version of this system that combines additional data and user input might help in scaling early detection.

Erraguntla et al. (2019), The Framework for Infectious Disease Analysis is a collection of standards for studying infectious illnesses. The effect of infectious illness on human populations is influenced by environmental factors, vector dynamics, transmission mechanics, social and cultural habits, and public policy. A comprehensive disease control paradigm must integrate the full disease lifecycle, including illness development from reservoir populations, zoonotic vector transmission, and effect on human society. The Framework for Infectious Disease Analysis is a software environment and conceptual framework for data integration, situational awareness, visualization, prediction, and intervention assessment. To analyze disease dynamics and test interventions in complex, heterogeneous populations, the Framework for Infectious Disease Analysis collects bio surveillance data using natural language processing, integrates structured and unstructured data from multiple sources, employs advanced machine learning, and employs multi-modeling. In the depicted case studies, natural language processing from social media, news feeds, and websites was used for information extraction, bio surveillance, and situation awareness. Classification machine learning approaches (support vector machines, random forests, and boosting) were used to predict illness.

Turjman et al. (2021), COVID-19, and other infectious disease identification on an AI-powered cloud. The COVID-19 pandemic has shown considerable gaps in our understanding of infectious disease transmission, prevention, and vaccine development and deployment. Some of these gaps might be filled by artificial intelligence, machine learning, big data, cloud computing, next-generation networks, and mobile devices. This special edition has 16 papers of outstanding quality. The first paper, ‘A systematic approach for COVID-19 predictions and parameter estimation’, attempted to predict and analyze the disease and its associated issues in order to determine the maximum number of infected people, the rate of spread, and, most importantly, the disease’s evaluation using a model-based parameter estimation method. In this work, COVID-19 was analysed using the Susceptible-Infectious-Recovered model under various situations. The impacts of lockout, the light switch mechanism, and parameter changes such as contact ratio and reproduction number are also studied. The authors sought to assess and anticipate the lockdown impact in terms of infected and recovered numbers, particularly in India, where there has been tremendous progress. The authors offer a disease-free endemic stability study utilizing Lyapunov and LaSalle’s techniques, as well as novel approaches such as the convalescent plasma method and the Who Acquires Infection From Whom method, both of which are believed to help flatten the COVID-19 curve.

3 Objective of the Study

- To Detect and predict of infectious disease.
- Perform automated learning to predict and accurately detect infection
- To construct the model that can perform in real time aspects

4 Research Methodologies

The data that gather from standard library separated into 70-30 ratios to train and verify the data collecting, as discussed above. A wide range of characteristics, including age, ethnicity, and other medical criteria, may be found in research used to make a diagnosis.

a. Modelling of data

During this step, the logical approach’s inputs are taken into account. The machine learning repository at the University of California, Irvine, collects data.

b. Treatment of Missing Values

All characteristics from the vector to the field of the device have been adjusted using noise cancellation and data normalisation as a priori model.

c. Data Analysis

A restricted number of machine learning approaches are used to alter the data gathered.

d. Construct the model

Last but not least, the data is used to model and build the components. Internet of Things sensors should be used in the custom framework to identify and predict infectious illnesses.

In order to arrive at a categorization rather than raw data, this thesis examines two instances of infectious illness. Pneumonia should be investigated first, and then the malaria illness categorization should be used. The high-accuracy ML has been merged with fuzzy logic in MATLAB. The fuzzy-based result has been provided by the layout that was created.

4.1 Flow Chart of the Proposed Work

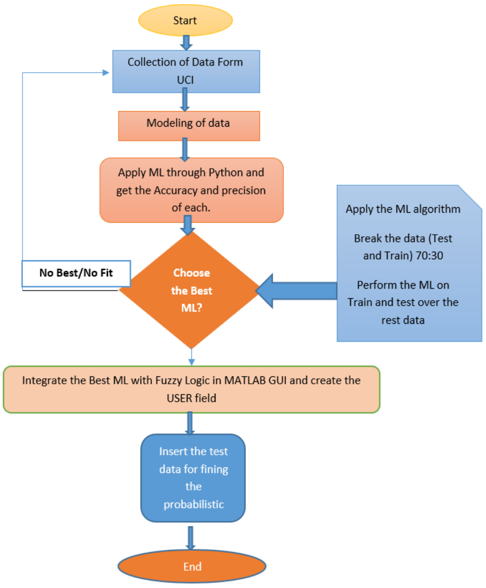

Start
Collection of Data Form UCI
Modeling of data
Apply ML through Python and get the Accuracy and precision of each.
Apply the ML algorithm
Break the data (Test and Train) 70:30
Perform the ML on Train and test over the rest data
No Best/No Fit
Choose the Best ML?
Integrate the Best ML with Fuzzy Logic in MATLAB GUI and create the USER field
Insert the test data for fining the probabilistic
End

Fig. 1: Flow Chart of the proposed work

4.2 Proposed Steps for Data Modelling

Phase 1: first and first, be certain that the data sets in question are, in fact, relevant. For statistical analysis, we choose the property with the least and biggest values in our dataset.

Phase 2: Analyses the data for mathematical patterns to see whether it is normal.

Phase 3: The 'Evaluate column mean' should be used to fill in the blanks where missing values should be entered.
Phase 4: The median and mean of the data sets should be used to fill in the missing values.
Phase 5: Use 70 percent of the data to train ML algorithms, and the other 30 percent for additional testing.
Phase 6: Analyse the train data using a machine learning technique.
Phase 7: The outcomes of the tests should be compared to the data sets used in the tests.

4.3 Methodology

Stage 1: Preparation of the data for further use or analysis.
A graphic depicting the information
Remove any outliers from the data set.
Classify and handle the missing information.
Using appropriate statistical methods to analyses the data.
The mean and median values should be used to fill in the blanks.

Stage 2: The process of picking a model.
{
Because of this, we're collecting data (classes)
Python-based application for machine learning algorithms.
}
Stage 3: in order to discover a better way of classifying the data
{
Input all the data that has been waiting using Python.
}
Stage 4: Accuracy is used to gauge the quality of the work.
{
Using the 'Performance' operator, you can determine how accurate your results are. Then use the 'accuracy measure' to determine the accuracy.
}
Stage 5: MATLAB's finest machine learning implementation
{
Use ML-Best to put the Fuzzy into action.
Set up the Insertion Text in the MATLAB GUI
Please provide the actual results of the exam.
}
Stage 6: Data Fuzziness should be tested.
The data is inserted and then executed.
Find the condition with a shaky probability
Pop-up a test result for the user.
}

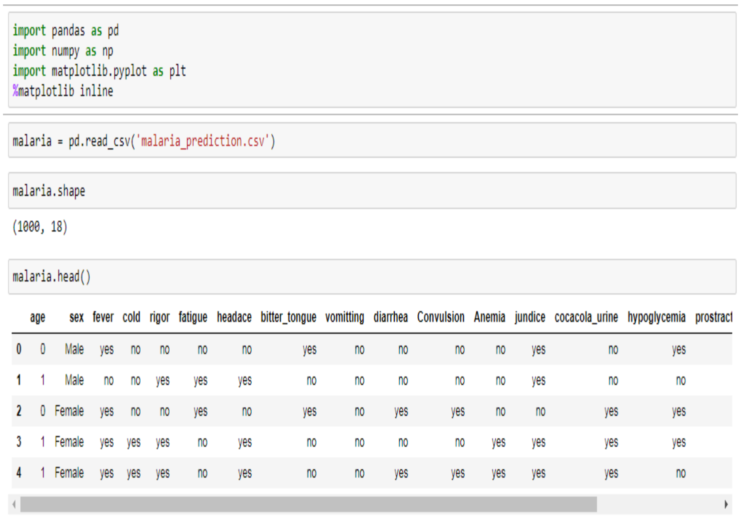
Pseudo Code

```
Let D={d1,d2,d3, … dn} be the given dataset
A= {}, the set of Algorithms classifiers
M= Mean and Median {M1, M2, M3, …Mn} for all Column,
Find the Missing Values
Replace M = Mean, Median of set
for ( i=vacant, i = 0, i++);
{
for (j= vacant, j = 0, j++);
}
Apply ML Algorithm
f = ML (Mod: Data);
Let D= {d1, d2,d3, … dn} be the given dataset
E = {E1, E2, E3, …En}, the set of ensemble classifiers
C = {c1, c2, c3, …cn}, the set of classifiers
X = the training set, X D
Y = the test set, Y D
K = meta level classifier
L = n(D)
for i =1 to L do
M(i) = Model trained using E(i) on X
Next i
M=M K
Result = Y classified by M
Main Result = Best {Result ¥ ML)
R []: Mix Fz with Max_ accuracy-ML
Test Y in R[i]
R[] = ∑R[i]
Fz = R[]
}
End
```

4.4 Implementation Phase

We have investigated on two diseases as one is Malaria Infectious Disease and second is the Covid. For malaria, we have explored the data and then perfrom the real time app to detection of Malaria Infectious through User data. Then we have do the same on Covid-19.

Impleemtation phase of Malaria Infectious Disease Perdiction



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

```
malaria = pd.read_csv('malaria_prediction.csv')
```

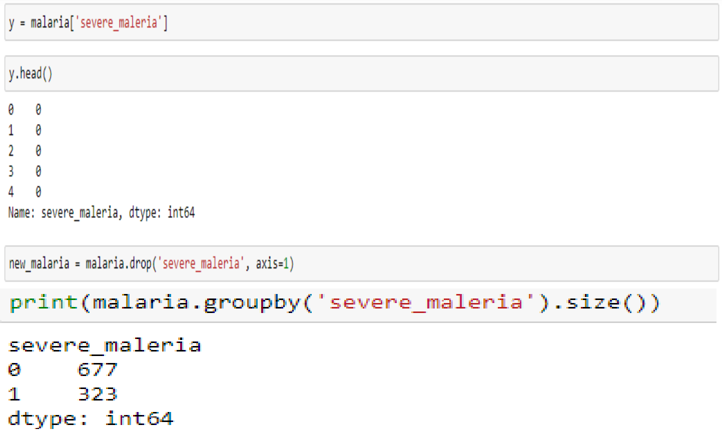
```
malaria.shape
```

```
(1000, 18)
```

```
malaria.head()
```

	age	sex	fever	cold	rigor	fatigue	headace	bitter_tongue	vomitting	diarrhea	Convulsion	Anemia	jundice	cocacola_urine	hypoglycemia	prostract
0	0	Male	yes	no	no	no	no	yes	no	no	no	no	yes	no	yes	
1	1	Male	no	no	yes	yes	yes	no	no	no	no	no	yes	no	no	
2	0	Female	yes	no	no	yes	no	yes	no	yes	yes	no	no	yes	yes	
3	1	Female	yes	yes	yes	no	yes	no	no	no	no	yes	yes	yes	yes	
4	1	Female	yes	yes	yes	no	yes	no	no	yes	yes	yes	yes	yes	no	

Fig. 2: Import libraries and read CSV file



```
y = malaria['severe_maleria']
```

```
y.head()
```

```
0    0
1    0
2    0
3    0
4    0
Name: severe_maleria, dtype: int64
```

```
new_malaria = malaria.drop('severe_maleria', axis=1)
```

```
print(malaria.groupby('severe_maleria').size())
```

```
severe_maleria
0    677
1    323
dtype: int64
```

Fig. 3: Severe Malaria

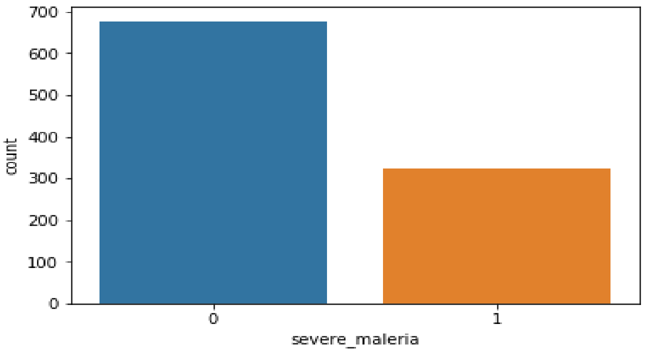


Fig. 4: Severe Malaria vs Count

K-Nearest Neighbors

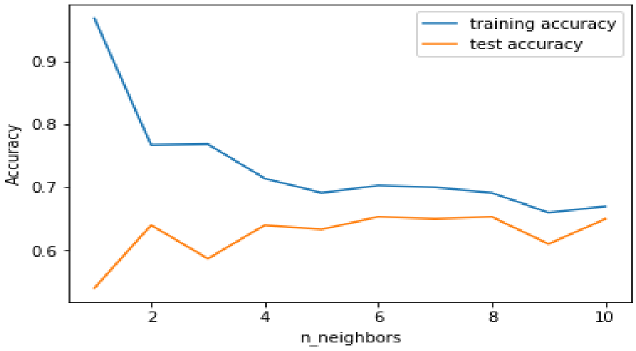


Fig. 5: K-Nearest Neighbors – neighbors Vs Accuracy

Accuracy of K-NN classifier on training set: 0.67
Accuracy of K-NN classifier on test set: 0.65

Logistic Regression
Training set accuracy: 0.674
Test set accuracy: 0.683

Decision Tree
Accuracy on training set: 0.969
Accuracy on test set: 0.607

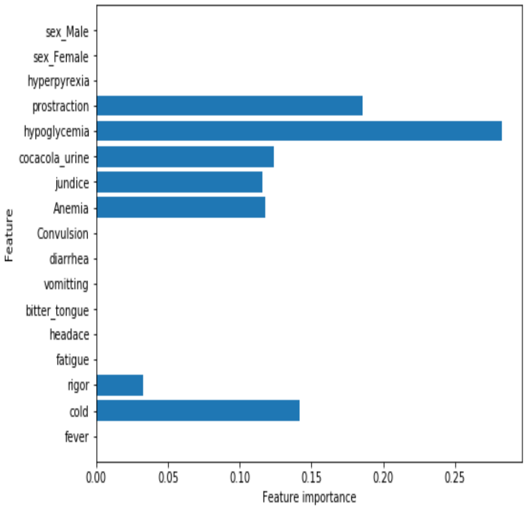


Fig. 6: Decision Tree – Feature Importance Vs Feature

Random Forest
Accuracy on training set: 0.674
Accuracy on test set: 0.683

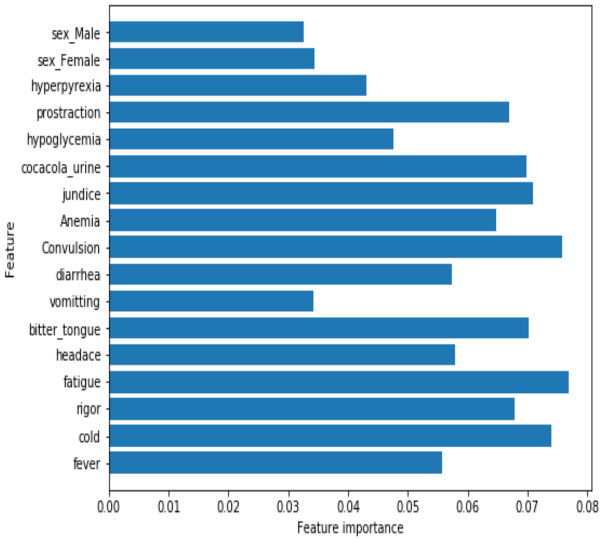


Fig. 7: Random Forest– Feature Importance Vs Feature

Gradient Boosting
Accuracy on training set: 0.717
Accuracy on test set: 0.623

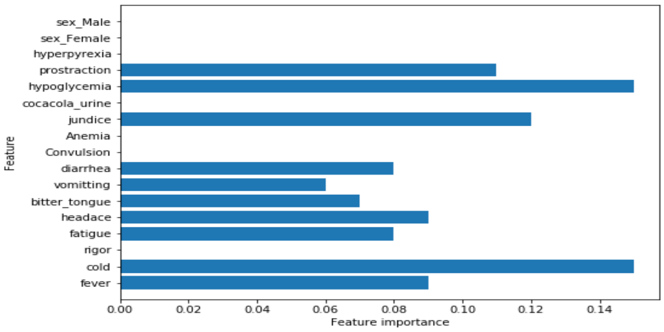


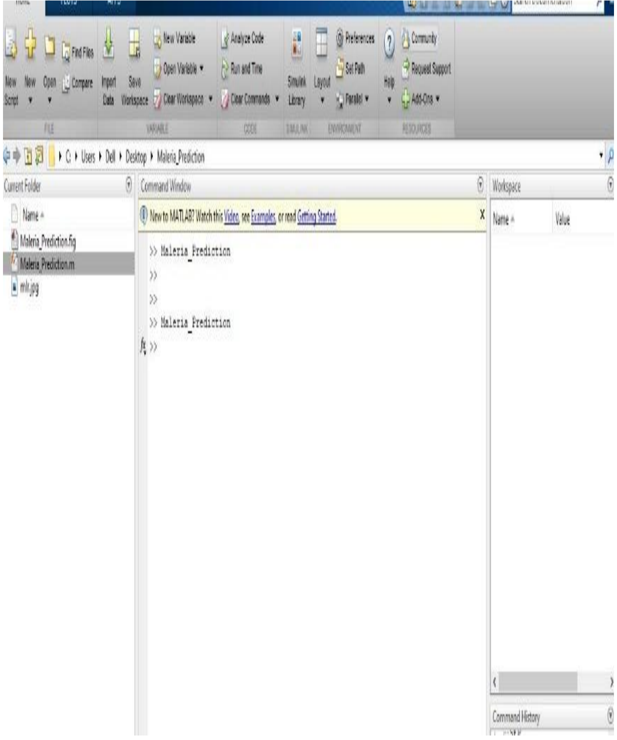
Fig. 8: Gradient Boosting– Feature Importance Vs Feature

Support Vector Machine
Accuracy on training set: 0.67
Accuracy on test set: 0.68

Table.1: Accuracy on training set and Accuracy on test set of Different Machine learning Methods

Methods	Accuracy on training set	Accuracy on test set
K-Nearest Neighbors	0.68	0.67
Logistic Regression	0.67	0.68
Decision Tree	**0.97**	0.61
Random Forest	0.67	0.68
Gradient Boosting	0.72	0.63
Support Vector Machine	0.67	0.68

Proposed MATLAB layout for predicting real-time patient status from minimal diagnostic data utilising a Decision Tree and Fuzzy logic classification. Extensive Fuzzy probabilistic model is the term given to the suggested system in this thesis. MATLAB has been used to implement the concept. The default layout has also been tested, as seen in the following image.



C: ▸ Users ▸ Dell ▸ Desktop ▸ Maleria_Prediction

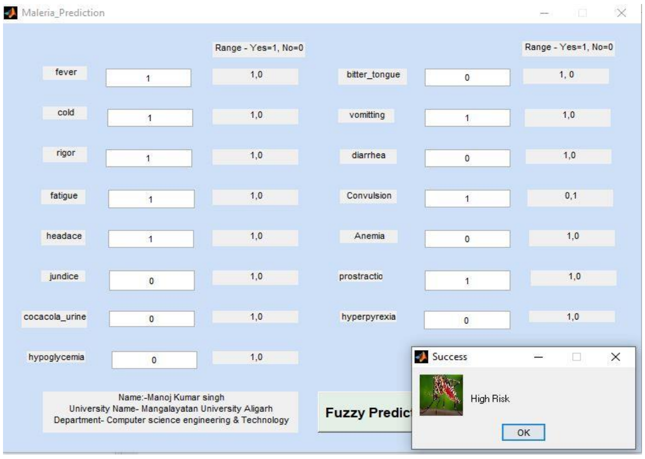
Maleria_Prediction.fig
Maleria_Prediction.m
mh.jpg

New to MATLAB? Watch this Video, see Examples, or read Getting Started.

```
>> Maleria_Prediction
>>
>>
>> Maleria_Prediction
fx >>
```

Fig. 9: Proposed Layout constructed in MATLAB through Decision Tree

The MATLAB program's default settings are shown above, and the intended work may be found in the Current folder using this method. It has been tested in command windows as a standalone executable file.



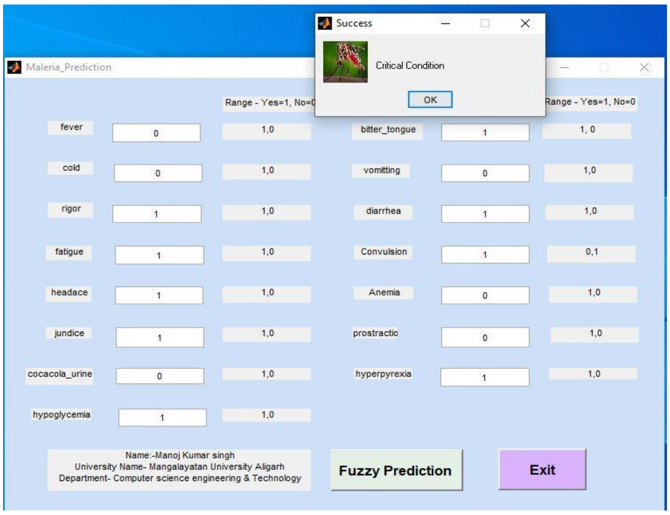
		Range - Yes=1, No=0			Range - Yes=1, No=0
fever	1	1,0	bitter_tongue	0	1, 0
cold	1	1,0	vomitting	1	1,0
rigor	1	1,0	diarrhea	0	1,0
fatigue	1	1,0	Convulsion	1	0,1
headace	1	1,0	Anemia	0	1,0
jundice	0	1,0	prostractio	1	1,0
cocacola_urine	0	1,0	hyperpyrexia	0	1,0
hypoglycemia	0	1,0			

Name:-Manoj Kumar singh
University Name- Mangalayatan University Aligarh
Department- Computer science engineering & Technology

Success
High Risk

Fig. 10: Proposed Layout constructed in MATLAB through Decision Tree-2

This is a straightforward representation of a mathematical map that has been created. The words 'fever', 'cold', and 'rigour' are all included in the preceding statement. As a result of this, MATLAB users often find themselves needing to bring in a new value while creating an outline.



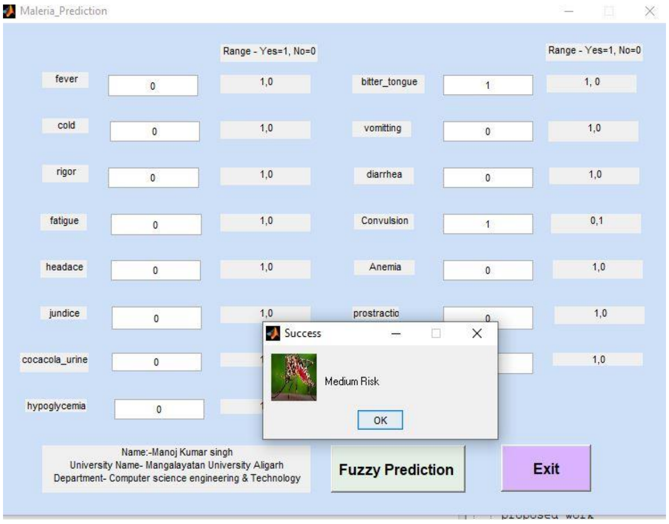
		Range - Yes=1, No=0			Range - Yes=1, No=0
fever	0	1,0	bitter_tongue	1	1, 0
cold	0	1,0	vomitting	0	1,0
rigor	1	1,0	diarrhea	1	1,0
fatigue	1	1,0	Convulsion	1	0,1
headace	1	1,0	Anemia	0	1,0
jundice	1	1,0	prostractio	0	1,0
cocacola_urine	0	1,0	hyperpyrexia	1	1,0
hypoglycemia	1	1,0			

Name:-Manoj Kumar singh
University Name- Mangalayatan University Aligarh
Department- Computer science engineering & Technology

Success
Critical Condition

Fig. 11: Fuzzy Prediction - Critical Condition

The 'blue-tinged' chamber contains the real data, as can be seen in the preceding screenshot. Finally, let's talk about how information is included into the design. In the popup box, it was noted that the result you obtained had the critical circumstance.



		Range - Yes=1, No=0			Range - Yes=1, No=0
fever	0	1,0	bitter_tongue	1	1, 0
cold	0	1,0	vomitting	0	1,0
rigor	0	1,0	diarrhea	0	1,0
fatigue	0	1,0	Convulsion	1	0,1
headace	0	1,0	Anemia	0	1,0
jundice	0	1,0	prostractio	0	1,0
cocacola_urine	0				1,0
hypoglycemia	0				

Name:-Manoj Kumar singh
University Name- Mangalayatan University Aligarh
Department- Computer science engineering & Technology

Success
Medium Risk

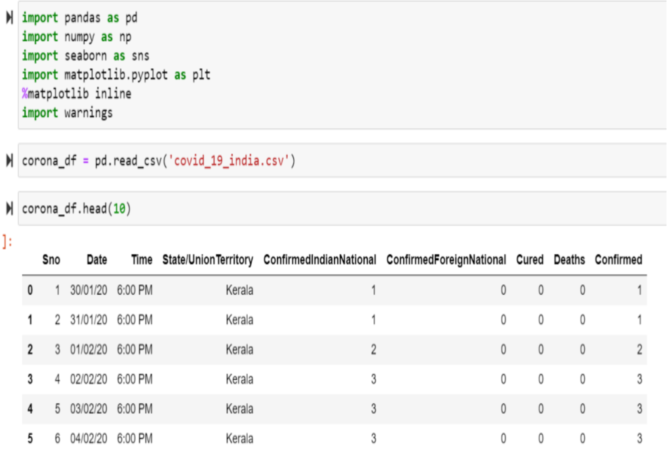
Fig. 12: Fuzzy Prediction – Medium Condition

When the operation was conducted, the outcome seemed to be at high risk in the popup. After putting in the patient's information, the above formula determines the medium risk. The approach used to discover a patient's cardiac details yielded results that were difficult to pin down.

4.5 Exploration and Investigation of covid-19

Due to Covid-19, there has been a dramatic loss in human life on the planet and this pandemic has presented an unparalleled threat to public health. In spite of the tremendous advances in wealth and innovation, about 50 million citizens are facing severe economic and social distress due to a catastrophic covid-19, with slow possibilities for recovery. The growth and spreading of this virus must be very concerning for every country. Various researches have been performed to know the spreading scenario of many countries has already been discussed in the reviews. Still very less research in Indian context with graphical exploration. This research explores the covid Spreading Scenario in India (Covid-19 Spreading

Scenario in India, which is the second most populous country in the world is suffering from the COVID-19 disease, which is known to be incurable. On or around May 18th, Indian authorities investigated approximately 0.1 million COVID-19-infected cases, and as of July 11th, these cases numbered 0.8 million. This analysis took a look at the corona from the most intense phase to the start of expansion.



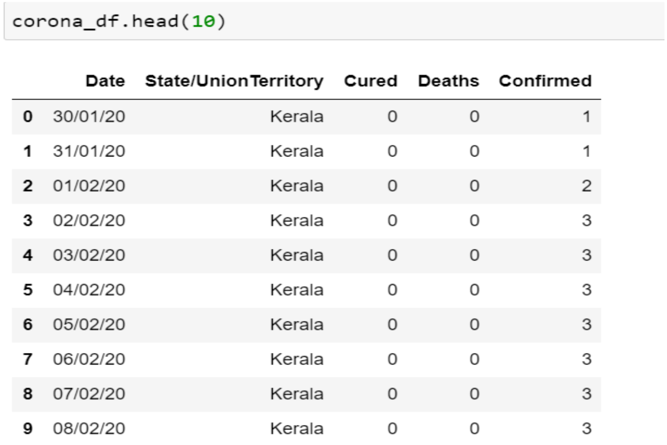
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
```

```
corona_df = pd.read_csv('covid_19_india.csv')
```

```
corona_df.head(10)
```

	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1	30/01/20	6:00 PM	Kerala	1	0	0	0	1
1	2	31/01/20	6:00 PM	Kerala	1	0	0	0	1
2	3	01/02/20	6:00 PM	Kerala	2	0	0	0	2
3	4	02/02/20	6:00 PM	Kerala	3	0	0	0	3
4	5	03/02/20	6:00 PM	Kerala	3	0	0	0	3
5	6	04/02/20	6:00 PM	Kerala	3	0	0	0	3

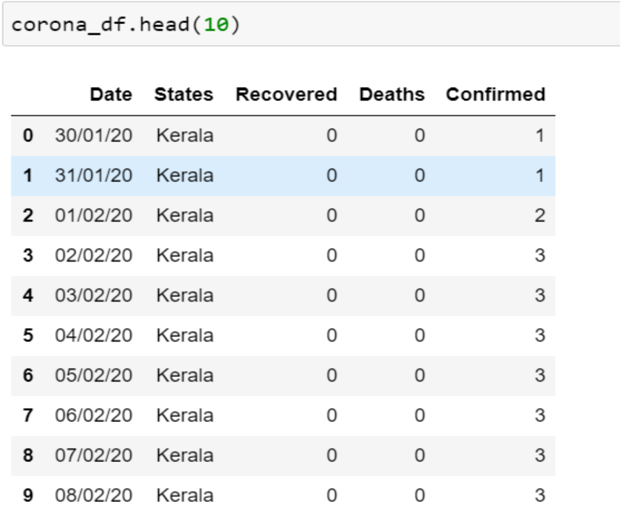
Fig. 13: Python Libraries and dataset



```
corona_df.head(10)
```

	Date	State/UnionTerritory	Cured	Deaths	Confirmed
0	30/01/20	Kerala	0	0	1
1	31/01/20	Kerala	0	0	1
2	01/02/20	Kerala	0	0	2
3	02/02/20	Kerala	0	0	3
4	03/02/20	Kerala	0	0	3
5	04/02/20	Kerala	0	0	3
6	05/02/20	Kerala	0	0	3
7	06/02/20	Kerala	0	0	3
8	07/02/20	Kerala	0	0	3
9	08/02/20	Kerala	0	0	3

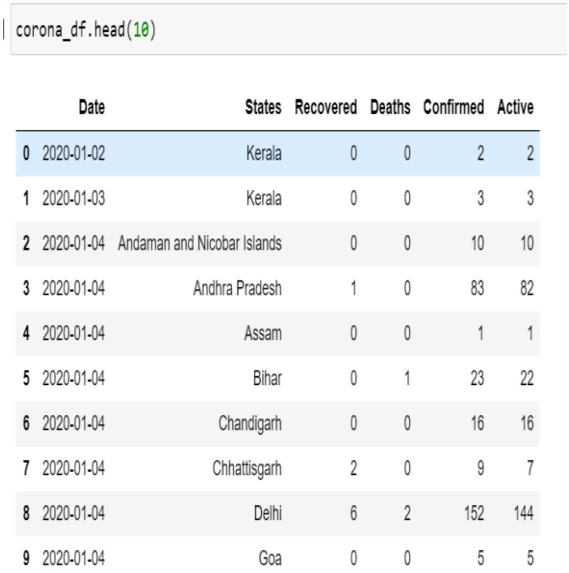
Fig. 14: Cured Deaths and Confirmed Case (A)



```
corona_df.head(10)
```

	Date	States	Recovered	Deaths	Confirmed
0	30/01/20	Kerala	0	0	1
1	31/01/20	Kerala	0	0	1
2	01/02/20	Kerala	0	0	2
3	02/02/20	Kerala	0	0	3
4	03/02/20	Kerala	0	0	3
5	04/02/20	Kerala	0	0	3
6	05/02/20	Kerala	0	0	3
7	06/02/20	Kerala	0	0	3
8	07/02/20	Kerala	0	0	3
9	08/02/20	Kerala	0	0	3

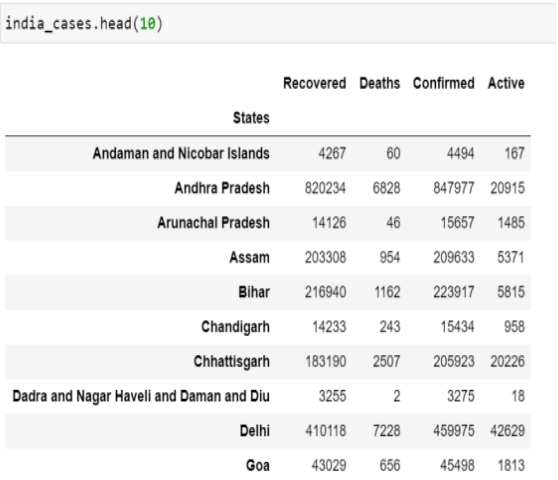
Fig. 15: Cured Deaths and Confirmed Case (B)



```
corona_df.head(10)
```

	Date	States	Recovered	Deaths	Confirmed	Active
0	2020-01-02	Kerala	0	0	2	2
1	2020-01-03	Kerala	0	0	3	3
2	2020-01-04	Andaman and Nicobar Islands	0	0	10	10
3	2020-01-04	Andhra Pradesh	1	0	83	82
4	2020-01-04	Assam	0	0	1	1
5	2020-01-04	Bihar	0	1	23	22
6	2020-01-04	Chandigarh	0	0	16	16
7	2020-01-04	Chhattisgarh	2	0	9	7
8	2020-01-04	Delhi	6	2	152	144
9	2020-01-04	Goa	0	0	5	5

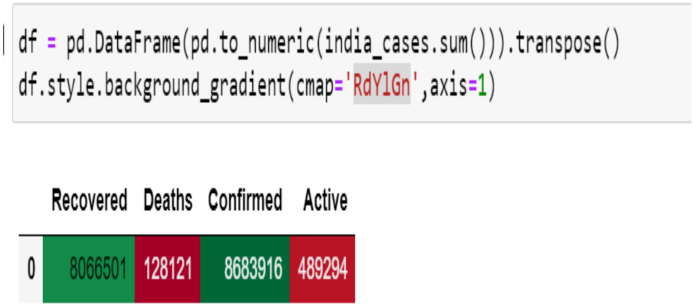
Fig. 16: Cured Deaths and Confirmed Case and Active (B)



```
india_cases.head(10)
```

	Recovered	Deaths	Confirmed	Active
States				
Andaman and Nicobar Islands	4267	60	4494	167
Andhra Pradesh	820234	6828	847977	20915
Arunachal Pradesh	14126	46	15657	1485
Assam	203308	954	209633	5371
Bihar	216940	1162	223917	5815
Chandigarh	14233	243	15434	958
Chhattisgarh	183190	2507	205923	20226
Dadra and Nagar Haveli and Daman and Diu	3255	2	3275	18
Delhi	410118	7228	459975	42629
Goa	43029	656	45498	1813

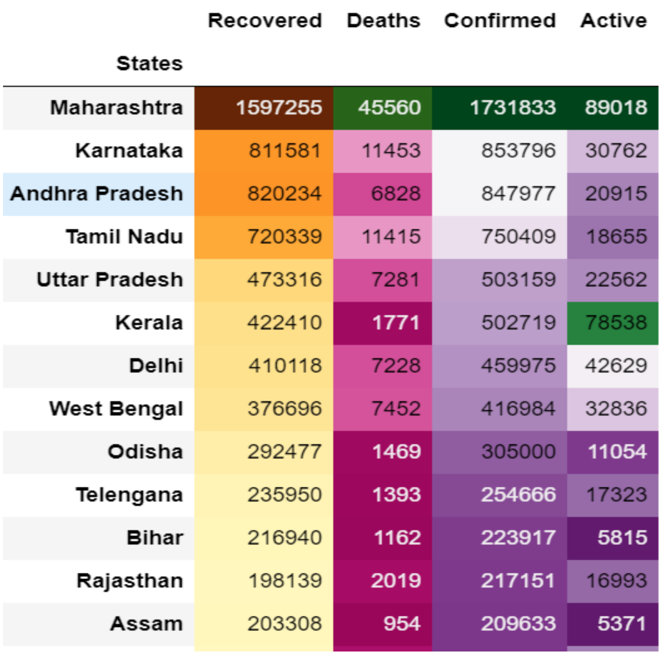
Fig. 17: Recovered Death Confirmed Active Case (A)



```
df = pd.DataFrame(pd.to_numeric(india_cases.sum())).transpose()
df.style.background_gradient(cmap='RdYlGn',axis=1)
```

	Recovered	Deaths	Confirmed	Active
0	8066501	128121	8683916	489294

Fig. 18: Total Recovered Deaths and Confirmed and Active Case



States	Recovered	Deaths	Confirmed	Active
Maharashtra	1597255	45560	1731833	89018
Karnataka	811581	11453	853796	30762
Andhra Pradesh	820234	6828	847977	20915
Tamil Nadu	720339	11415	750409	18655
Uttar Pradesh	473316	7281	503159	22562
Kerala	422410	1771	502719	78538
Delhi	410118	7228	459975	42629
West Bengal	376696	7452	416984	32836
Odisha	292477	1469	305000	11054
Telengana	235950	1393	254666	17323
Bihar	216940	1162	223917	5815
Rajasthan	198139	2019	217151	16993
Assam	203308	954	209633	5371

Fig. 19: Total Recovered Deaths and Confirmed and Active Case in states

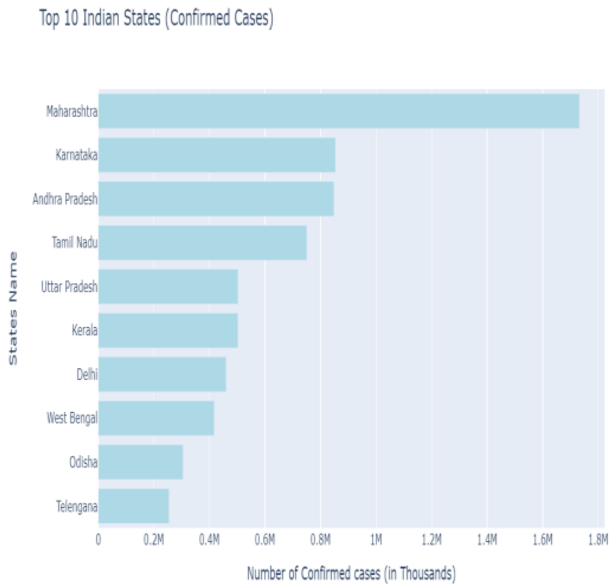


Fig. 20: Top 10 Indian States Confirmed cases

As shown in Figure the Covid scenario of top 10 India states (Confirm Case) presented in bar chart. The highest found in Maharashtra.

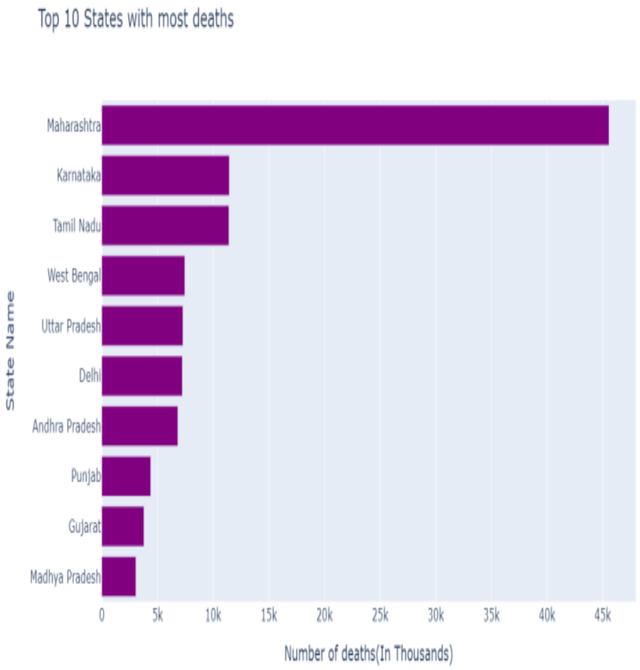


Fig. 21: Top 10 States with Most deaths

As the above figure suggested that the Covid scenario of top 10 Most Death presented in bar chart. The highest found in Maharashtra.

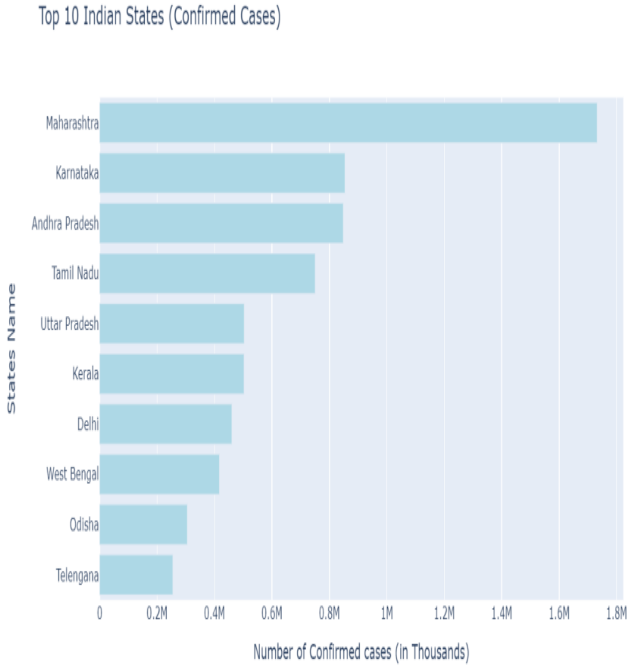


Fig. 22: Top 10 States with confirmed cases

As the above figure suggested that the Covid scenario of top 10 India states presented in bar chart. The highest found in Maharashtra.

.

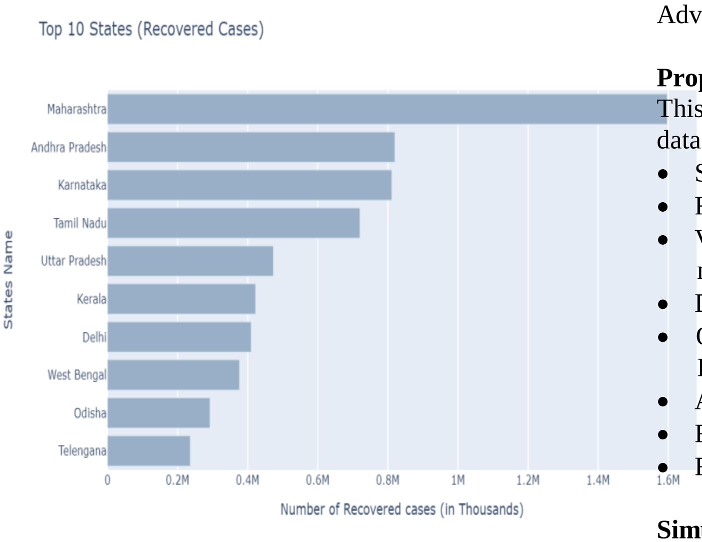


Fig. 23: Top 10 States with Recovered Cases

As the above figure suggested that the Covid scenario of top 10 states (in Recovery) presented in bar chart. The highest found in Maharashtra.

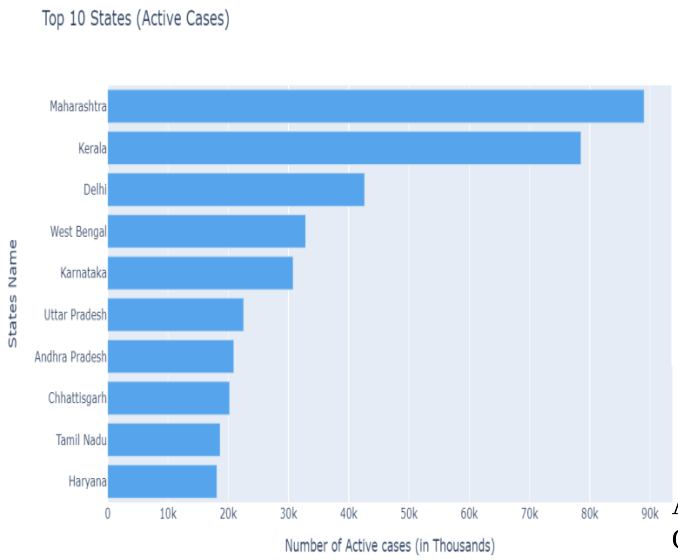


Fig. 24: Top 10 States with Active cases

As the above figure suggested that the Covid scenario of top 10 states (Active State) presented in bar chart. The highest found in Maharashtra. The methodology of this research work can be described in following steps:

Step 1: Current research the infection rate and mortality rate of Covid.

Step 2: To extract and analyses relevant data from the rate of infection and death of Covid.

Step 3: Apply techniques of regression analysis to estimate covid infection rate and death rate

Step 4: To find the forecast for the next five years with a regression and line forecast analysis of covid infection and death rate.

Step 5: To collect data from recent survey-based research to obtain an understanding of covetousness in Indian contexts.

Tools Used

Adv Excel Tools (Data Analytic Tool Packs)

Proposed Framework for covid prediction

This prediction involves the following step through which the data has been finally modelled for prediction.

- Shorting the most significant data.
- Find the trends analysis of data through ROC curve.
- Value therapies that are not being used (Replace Mean or median values in vacant spaces)
- Divide the data sets in two parts.
- One set is for testing, while the second is for training Purposes (Prefer 70:30 ration for train: test)
- Apply the regression model over the data sets
- Find the accuracy with precision
- Find the best accuracy algorithm.

Simulation Result for prediction of covid

The exploration and analysis of the covid data set performed through MS Excel and analytic tools packs. As the figure below the confirmed case of covid has been presented of different across the world.

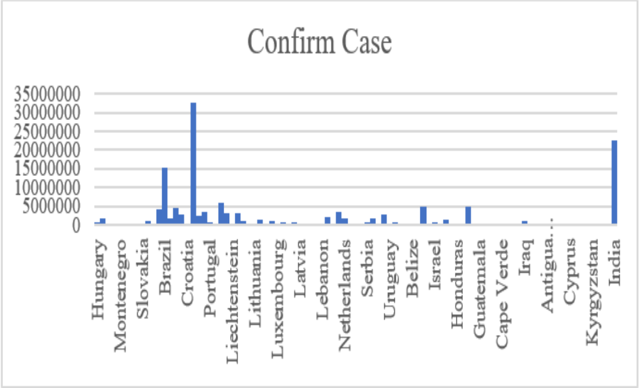


Fig. 25: Covid-19 cases and Confirm Case by country as of 10 May 2021

As the above figure illustrate that the India status of Covid -19 scenario world wise cases. The above bar chart represented the confirm case of Covid. India has been at second position.

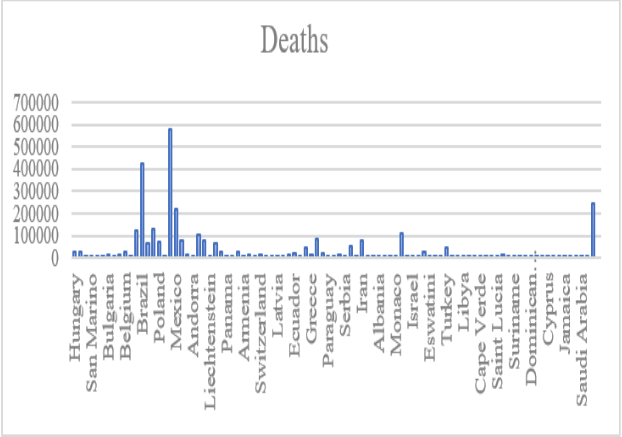


Fig. 26: Covid -19 scenario world wise cases

As the above figure illustrate that the India status of Covid -19 scenario world wise cases. The above bar chart represented the Death of Covid. India has been at third position.Covid-19 cases and Deaths by country as of 10 May 2021.

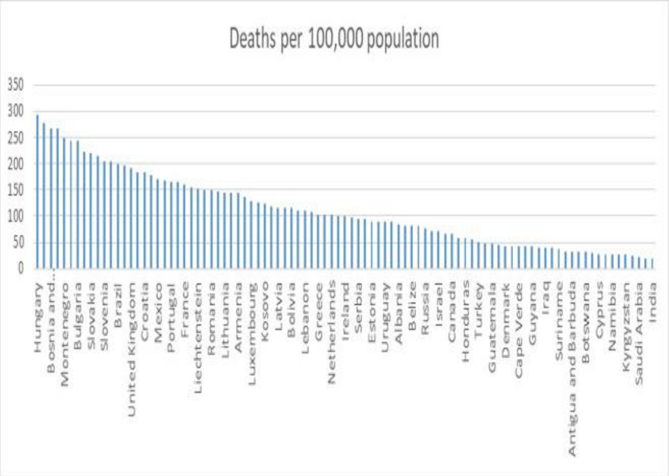


Fig. 27: Covid-19 cases and Deaths per 10000 by country as of 10 May 2021

As the above figure illustrate that the India status of Covid -19 scenario world wise cases. The above bar chart represented the Death per one lakhs population of Covid. India has been found significantly very a smaller number of deaths.

Forecasting of Covid

The time series data of covid has been apply over the trend analysis through Adv excel.

Table 2: Forecasting of Covid

Time	Day	Total Case
03/10	1	47.00
03/11	2	60.00
03/12	3	74.00
03/13	4	81.00
03/14	5	84.00
03/15	6	110.00
03/16	7	114.00
03/17	8	137.00
03/18	9	150.00
03/19	10	171.00
03/20	11	223.00
03/21	12	283.00
03/22	13	360.00
03/23	14	434.00
03/24	15	519.00
03/25	16	606.00

The data has been continuing from 10 March 2020 to 06 May 2021. Total number of occurrences of covid case has been represented correspondingly. The graph has further forecast from June 2021 to June 2022. The future one-year prediction has been drawn in yellow line. Whereas the current data reflected by blue line. The figure below presented in million in y axis whereas the x axis is time in months.

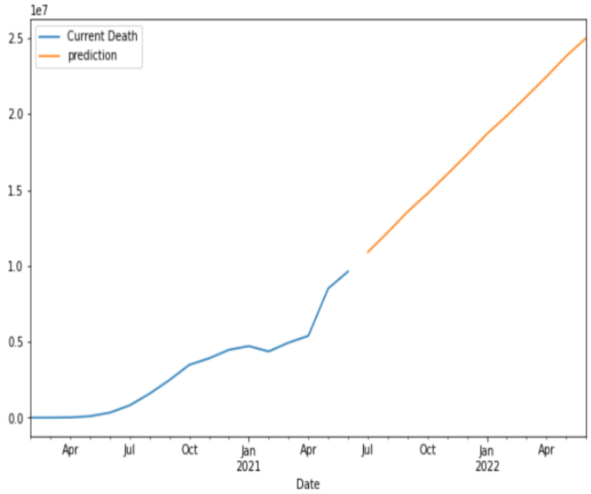


Fig. 28: Covid-19 Prediction Chart

So, the trend line reflected that the covid case in next year reaches more than 25 lakhs cases. This model used **Holt-Winters's** method which is **a highly acceptable time series forecasting method. This method is capable of forecasting the data.**

Table 3: Accuracy on training set and Accuracy on test set of Different Machine learning Methods (Covid- diseases)

Methods	Accuracy on training set	Accuracy on test set
K-Nearest Neighbors	0.90	0.89
Logistic Regression	0.89	0.87
Decision Tree	**0.93**	0.78
Neural Network	0.67	0.68

Table 4: Comparison of Existing model and Proposed model

	Existing model Cheshmehzangi et.al., (2021)	**Proposed model**
1. Platform	SEIR Data Analysis	Python
2. Base Research	Data Analysis and Exploration	Data Analysis, Exploration, Forecasting
3. Research Area	Covid-19	Covid-19
4. Method	SEIR Model	**Holt-Winters's**
5. Modules	Single (only Exploration)	Two (Exploration+ Prediction Method)
6. Outcome	In Graphical Form	Graphical and Numeric

As the proposed scenario has apply Data Analysis, Exploration, forecasting whereas the existing research has only exploration based on Data Analysis. The proposed feature profound the prediction and well as the analytical method to this investigation.

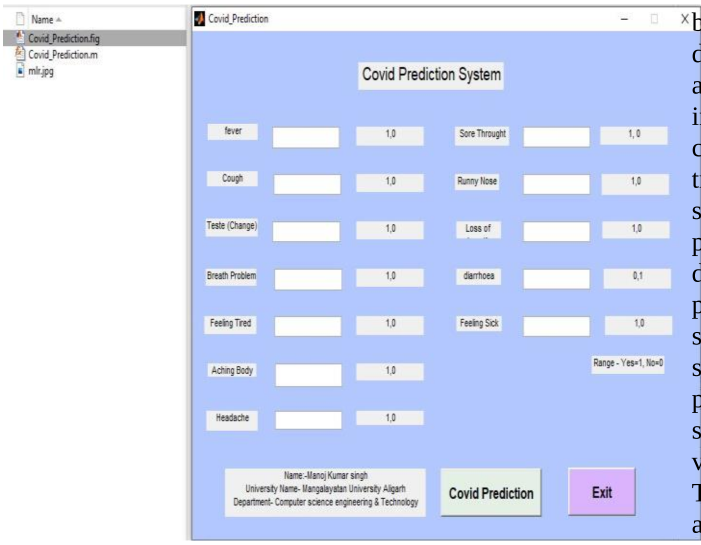
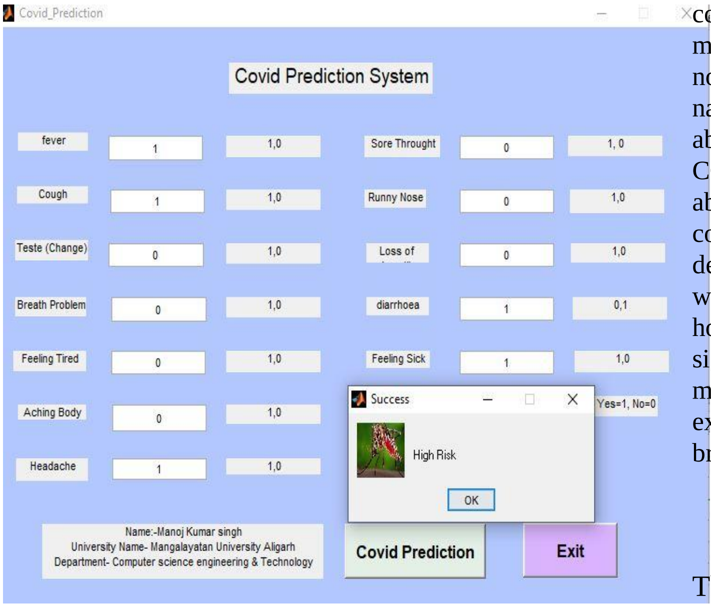


Fig. 29: Real time user interface for Covid

This is default layout constructed for user interface to check the covid.


Covid_Prediction

Covid Prediction System

fever 1 1,0
Cough 1 1,0
Teste (Change) 0 1,0
Breath Problem 0 1,0
Feeling Tired 0 1,0
Aching Body 0 1,0
Headache 1 1,0

Sore Throught 0 1, 0
Runny Nose 0 1,0
Loss of 0 1,0
diarrhoea 1 0,1
Feeling Sick 1 1,0
Yes=1, No=0

Success
High Risk
OK

Name:-Manoj Kumar singh
University Name- Mangalayatan University Aligarh
Department- Computer science engineering & Technology

Covid Prediction
Exit

Fig. 30: Covid Prediction System

5 CONCLUSION & FUTURE SCOPE

Infectious disease treatment takes the early identification of a patient's viruses and steps, such as vaccines to deter and/or heal them. In addition to detecting the more serious diseases that internists are capable of healing, internists are capable of catching the pathogens that are difficult to treat. There are several benefits to learning about infectious diseases such as fungi, bacteria, and viruses, including gaining a better knowledge of human pathogens and the need of medications and vaccinations. The straightforward epidemiological concepts and environmental, occupational, and host variables that generate susceptibility are not a problem for physicians who specialise in this subject. In the medical field, there are a variety of quality control measures. These crucial actions may be carried out by infectious disease clinicians (disease doctors) with a particular certification or with exceptional accuracy in a single clinic. All but a small percentage of infectious disease physicians are consultants to other health care providers and also accompany patients while they get treatment for their illnesses. Several infectious disease specialists have access to adequate medical facilities to practise their craft. General health services are available to doctors who specialise in infectious diseases while treating patients recuperating from a variety of illnesses. Medical specialists that specialise in infectious diseases conduct scientific and clinical reviews, train medical students, and provide ongoing consultation to patients and surgical staff.This studies also gathered more evidence, with a larger variety of data sources, in the name of public health security. This research discovered that one of the difficulties in our analysis was the need to grasp and implement visualizations. This research also discovered that consumer trust and organization assistance played a major part in the assimilation of these approaches. We have now been quite serious with interoperability. When discussing a new product, one may solve difficulties with the usage of other tools, for example, by describing items that might be overlooked or that may communicate confusing knowledge. One of the popular models used to predict the COVID-19 pandemic's spread is now mainstream culture. It is impossible to discern which narrative is right since each paradigm offers a different story about the imminent extinction of existence. There are COVID-19 variants, but the other versions are not meant to be absolute facts. Rather of trying to predict the future, they consider a number of different possibilities that may arise depending on existing conditions. In the expectation that we would make them more accessible, as well as emphasizing how assumptions behind the models may result in significantly different estimations, they have suggested making projections public. Predictions like these let us examine the best- and worst-case situations, while also bringing us closer to our best-case scenario.

6 DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://www.kaggle.com/datasets/anandhuh/latest-covid19-india-statewise-data

7 ADVANTAGES OF SOCIETY

Our research on the detection and prediction of infectious diseases through data mining and machine learning has several potential advantages for society:

1. **Improved public health:** By using data mining and machine learning to better understand and predict the spread of infectious diseases, public health

officials can develop more effective strategies for controlling outbreaks and protecting the public.

2. **Faster and more accurate diagnosis:** By using machine learning algorithms to analyze data from multiple sources, healthcare providers can quickly and accurately diagnose infectious diseases, leading to better treatment outcomes and reduced transmission rates.
3. **More targeted interventions:** By identifying high-risk populations and transmission routes, public health officials can more effectively target their interventions and allocate resources where they are needed most.
4. **Faster development of treatments and vaccines:** By using machine learning to analyze medical research data, researchers can identify potential drug candidates more quickly, accelerating the development of new treatments and vaccines.
5. **Economic benefits:** Controlling the spread of infectious diseases can have significant economic benefits, including reducing healthcare costs, improving worker productivity, and preventing disruptions to supply chains and other critical infrastructure.

Overall, Our research has the potential to improve public health, accelerate the development of new treatments and vaccines, and have significant economic benefits for society.

The detection and prediction of infectious diseases (specifically COVID-19) through data mining and machine learning has a wide range of future scope. Here are some potential areas of future research:

i. **Expand to other infectious diseases:** The methods and techniques we developed for COVID-19 detection and prediction can be applied to other infectious diseases. By expanding your research to other diseases, we can help public health officials better understand and control the spread of a wide range of infectious diseases.

ii. **Integration with other data sources:** We can expand your research by integrating data from other sources, such as social media and remote sensors, to further improve the accuracy and timeliness of your predictions.

iii. **Advanced data analysis techniques:** We can explore advanced data analysis techniques, such as deep learning, to improve the accuracy and speed of your predictions.

iv. **Real-time monitoring and alerting:** We can develop a real-time monitoring and alerting system that can provide early warning of potential outbreaks, enabling public health officials to take swift action to control the spread of disease.

v. **International collaboration:** We can collaborate with researchers and healthcare professionals from other countries to develop a global infectious disease monitoring and prediction system, which can help prevent the spread of infectious diseases across borders.

vi. **Application to non-healthcare fields:** The techniques we developed can also be applied to other fields, such as finance and supply chain management, to predict and manage risks.

Overall, your thesis work has significant potential for future research and application in a variety of fields. By continuing to refine and expand your research, we can contribute to the development of more effective strategies for controlling the spread of infectious diseases, improving public health, and enhancing our understanding of disease transmission dynamics.

8 NO CONFLICT OF INTEREST STATEMENT

On behalf of all authors, the corresponding author states that there is no conflict of interest.

9 REFERENCES

1.Bloom, D. E., &Cadarette, D. (2019). Infectious disease threats in the twenty-first century: strengthening the global response. *Frontiers in immunology*, *10*, 549.

2.Agrebi, S., &Larbi, A. (2020). Use of artificial intelligence in infectious diseases. In *Artificial intelligence in precision health* (pp. 415-438). Academic Press.

3.Morse, S. S. (2001). Factors in the emergence of infectious diseases. *Plagues and politics*, 8-26.

4.Cohen, R., Ashman, M., Taha, M. K., Varon, E., Angoulvant, F., Levy, C., ...&Grimprel, E. (2021). Pediatric Infectious Disease Group (GPIP) position paper on the immune debt of the COVID-19 pandemic in childhood, how can we fill the immunity gap?. *Infectious Diseases Now*, *51*(5), 418-423.

5.Ortiz-Prado, E., Simbaña-Rivera, K., Gomez-Barreno, L., Rubio-Neira, M., Guaman, L. P., Kyriakidis, N. C., ...& Lopez-Cortes, A. (2020). Clinical, molecular, and epidemiological characterization of the SARS-CoV-2 virus and the Coronavirus Disease 2019 (COVID-19), a comprehensive literature review. *Diagnostic microbiology and infectious disease*, *98*(1), 115094.

6.Sohrabi, C., Alsafi, Z., O'neill, N., Khan, M., Kerwan, A., Al-Jabir, A., ...& Agha, R. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International journal of surgery*, *76*, 71-76.

7.Tran, H. N., Le, G. T., Nguyen, D. T., Juang, R. S., Rinklebe, J., Bhatnagar, A., ...& Chao, H. P. (2021). SARS-CoV-2 coronavirus in water and wastewater: A critical review about presence and concern. *Environmental research*, *193*, 110265.

8.Stadnytskyi, V., Anfinrud, P., &Bax, A. (2021). Breathing, speaking, coughing or sneezing: What drives transmission of SARS-CoV-2?. *Journal of Internal Medicine*, *290*(5), 1010-1027.

9.UmaMaheswaran, S. K., Munagala, N. K., Mishra, D., Othman, B., SINTHU, S., &Tripathi, V. (2022, April). The role of implementing Machine Learning approaches in enhancing the effectiveness of HealthCare service. In *2022*

2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1922-1926). IEEE.

10.Sankar, P., Ahmed, W. N., Koshy, V. M., Jacob, R., &Sasidharan, S. (2020). Effects of COVID-19 lockdown on type 2 diabetes, lifestyle and psychosocial health: a hospital-based cross-sectional survey from South India. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(6), 1815-1819.

11.Shi, Y., Wang, G., Cai, X. P., Deng, J. W., Zheng, L., Zhu, H. H., ... & Chen, Z. (2020). An overview of COVID-19. *Journal of Zhejiang University-SCIENCE B*, *21*(5), 343-360.

12.Turjman, F.A. (2021). AI-powered cloud for COVID-19 and other infectious disease diagnosis. *Personal and Ubiquitous Computing,* s00779-021.

13.Erraguntla, M., Zapletal, J., &Lawley, M. (2019). Framework for Infectious Disease Analysis. *Health Informatics Journal,* 25(4), 1170–1187.

14.Bogu, G.K., & Snyder, M.P. (2021). Deep learning-based detection of COVID-19 using wearables data. *medRxiv preprint,* 212-49474.

15.Wu, J., Zhang, P., Zhang, L., Meng, W., & Li, J. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. *medRxiv preprint,* 2005-1136.

16.Khan, F.A., Rakhami, M.A., & Ahmad, S. (2021), Detection and Prediction of Diabetes Using Data Mining. 4(12), 3059-343.

17.Ravi, M.A., Gopal, M.V., & Roselyn, J.P. (2021), Detection of Infectious Disease using Non-Invasive Logistic Regression Technique. *IEEE Xplore,* 57(17).

18.Bruyndonckx, R., &Coenen, S. (2021). Analysing the trend over time of antibiotic consumption in the community: a tutorial on the detection of common change-points. *J AntimicrobChemother,* 7(6), ii79–ii85.

19.Martinez, J.T., & Pérez, M.G. (2021), A Novel Machine Learning-Based Approach for the Detection of SSH Botnet Infection. *Future Generation Computer Systems,* 10-1016.

20.Huang, S., & Yang, G. (2021), Artificial intelligence in the diagnosis of COVID-19. *International Journal of Biological Sciences,* 17(6), 1581-1587.

21.Kim, M., &Chae, K. (2020). Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches. *International Journal of Environmental Research and Public Health,* 2(13).

22.Pundir, A., &Bhardwaj, S. (2021). Prediction and Detection of Covid-19 Using Ml / Ai. *International Journal of Multidisciplinary and Current Educational Research (IJMCER),* 3(3), 2581-7027.

23.Nikooghadam, M., &Ghazikhani, A. (2020). COVID-19 Prediction Classifier Model Using Hybrid Algorithms in Data Mining. *Int J Pediatr,* 9(1), 12723-12737.

24.Borkenhagen,L.K., & Allen, M.W. (2021). Influenza virus genotype to phenotype predictions through machine learning. *Emerging Microbes & Infections,* 10(1), 1896-1907.

25.Arji, G., &Ahmadi, H. (2019). Fuzzy logic approach for infectious disease diagnosis. Biocybernetics and Biomedical Engineering.

26.Smith, K.P., & Kirby, E.J. (2020), Image analysis and artificial intelligence in infectious disease diagnostics. *Clinical Microbiology and Infection,* 6(13).

27.Amar, L.A., Taha, A. A., & Mohamed, M.Y. (2020). Prediction of the final size for COVID-19 epidemic using machine learning. *Infectious Disease Modelling,* 622e634.

28.Nilashi, M., Ahmadi, H., &Shahmoradi, L. (2018). A predictive method for hepatitis disease diagnosis using ensembles of neuro fuzzy technique. *Journal of Infection and Public Health.*

29.Salehi, A.W., Baglat, P., & Gupta, G. (2020). Review on Machine and Deep Learning Models for the Detection and Prediction of Coronavirus. *Journal Pre-proofs.*

30.Li, M. (2019). Study on the Grouping of Patients with Chronic Infectious Diseases Based on Data Mining. *Journal of Biosciences and Medicines,* 119-135.

31.Cruz, A.P.D., &Tumibay, G. M. (2019). Predicting Tuberculosis Treatment Relapse. *Journal of Computer and Communications,* 243-251.

32.Callejon-Leblic, M.A., Moreno-Luna, R., &Cuvillo, A.D. (2021). Loss of Smell and Taste Can .Accurately Predict COVID-19 Infection. *Clin. Med,* 1004-0570

33.Kivrak, M., &Colak, C. (2021). Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. *Computer Methods and Programs in Biomedicine,* 20(1), 105951.

34.Kukar, M., &Gncer, G. (2020). COVID‑19 diagnosis by routine blood tests using machine learning. *Scientific Reports,* 1598-021.

35.Lalmuanawma, S., &Hussain, J. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic. *Nonlinear Science, and Nonequilibrium and Complex Phenomena,* 110059.

36.Zeng, D., Cao, D., & Neill, D.B. (2021). Artificial intelligence enabled public health surveillance—from local detection to global epidemic monitoring and control. *Artificial Intelligence in Medicine,* 821259-2.00022-3.

37.Natarajan, Y., Kannan, S., &Mohanty, S.N. (2021). Survey of Various Statistical Numerical and Machine Learning Ontological Models on Infectious Disease Ontology. *Infectious Disease Ontology Computation,* 431–442.

38.Woldaregay, A.Z., &Launonen, L.K. (2020), A Novel Approach for Continuous Health Status Monitoring and Automatic Detection of Infection Incidences in People With Type 1 Diabetes Using Machine Learning Algorithms. *J Med Internet Res,* 22(8), e1891233

39. Heba M. Afify, Muhammad Zanaty, "Computational Predictions for Protein Sequences of COVID-19 Virus via Machine Learning Algorithms", Medical & Biological Engineering & Computing, Springer, vol.59, no.9, 2021.

40. Heba M. Afify, Muhammad Zanaty, "A Comparative Study of Protein Sequences Classification-Based Machine Learning Methods for COVID-19 Virus against HIV-1", Applied Artificial Intelligence, Taylor & Francis, vol. 35, no. 15, pp: 1733-1745, 2021.

41. Kamel K. Mohammed, Heba M. Afify, Ashraf Darwish, Aboul Ella Hassanien, "Automatic Scoring and Grading of COVID-19 Lung Infection Approach", Digital Transformation and Emerging Technologies for Fighting

COVID-19 Pandemic: Innovative Approaches, Studies in Systems, Decision and Control 322,pp:63-79, Springer ,March 2021.