An Intelligent System for Prediction of COVID-19 Case using Machine Learning Framework-Logistic Regression

Annwesha Banerjee Majumder¹, Somsubhra Gupta², Dharmpal Singh³ and Sourav Majumder⁴

¹Dept. of Information Technology, JIS College of Engineering, Kalyani, Nadia, West Bengal, India
²Amity University, Kolkata, West Bengal, India
³Dept. of Computer Science and Engineering, JIS College of Engineering, Kalyani, Nadia, West Bengal, India
⁴Wipro Technologies, Kolkata, West Bengal, India

Email Id: annwesha.banerjee@jiscollege.ac.in

Abstract: World is now at a critical condition. Covid-19 is the disease caused by corona virus newly identified [1], the pandemic affecting the human being a lot. As on date the total active cases in India are 3.23 Million and death cases are 59,449. Many scientists and medical practitioners are working hard to fight against this, in search of proper medicine and vaccine. Research is also going on in the field of machine learning and AI to predict the spread of disease and also in identification of the presence of the virus in human body, which will help the field of medical science. In this paper we have proposed a method to identify whether a patient has risk of COVID-19 using Logistic Regression model, considering multiple symptoms.

Keywords: COVID-19, Logistic regression, Machine Learning, AI

1. Introduction:

Till date COVID-19 is one of the most critical threats to human being. Many researchers across the globe are working hard for a solution to defeat the tiny invisible threat. In parallel to the medical science computer science is also contributing a lot in the COVID-19 research area by making critical data analysis faster. Different machine learning based classifier and predictive models have also been proposed which are helpful in prediction of the disease at an early stage. In this paper we have proposed a machine learning based model using Logistic regression for COVID-19 patient classification. The dataset used for the model has been collected from Kaggle and we are achieving 92% accuracy.

COVID-19:Corona virus is a large family of virus causes infectious disease. People infected by this virus would have mild to moderate respiratory illness. Respiratory symptoms, fever, cough, shortness of breath and breathing difficulties are reported as most common symptoms for COVID-19. Sometimes this infection creates pneumonia, severe acute respiratory syndrome, kidney failure and even death.[3]. It has also been observed that if patient has other disease like hypertension, heart diseases diabetes, renal infection, anxiety and depression then it leads to a more critical condition and needs instant medical support. As per the report published by WHO, till 23rd of August 2020 over 1.7 million new COVID-19 cases and 39 000 new deaths. [4]
Logistic Regression:

Logistic Regression is a machine learning approach, which can be used for classification problem. In a paper published in Proceedings of Belgian Royal Academy Pierre Francais Verhulst described logistic function and its properties by defining three parameters and the curve passing though these.[5]. It is very simple and one of the mostly used machine learning algorithms. Logistic regression is a statistical model for predicting binary classes. Maximum likelihood Estimation model is used in logistic regression. The dependent variable in here follows Bernoulli distribution. Logistic function or sigmoid function is a ‘S’ shaped curve which takes value between 0 and 1. 1 will be predicted if the curve goes to positive infinity and 0 if it goes to negative infinity. Figure 1 is a representation of a logistic function.

\[
\text{(1)}
\]

The above equation is the representation of a logistic regression.

![Figure 1: Representation of Logistic Regression Model][6]

2. Literature Survey

Machine learning plays a major role in disease prediction and analysis. Though COVID-19 is a new disease still works have done in COVID-19 prediction using machine learning. Continuous contribution is going on in this field of research.

Ahmed Hamad et.al has proposed an accurate classification model for COVID-19 using KNN variant. They have used the dataset collected from Italian society of medical and intervention radiology society. A clear comparative analysis has given among Modified KNN (MKNN), KNN for imperfect data (KNNimp) and
cost sensitive KNN(csKNN) and the KNN variant has achieved most efficient performance.[7]. In another study [8] authors have collected blood samples from 404 patient in Wuhan China for identification of predictive biomarkers of the disease. An artificial intelligence based COVID-19 mortality prediction model has been proposed by authors. They have used Artificial Neural Networks, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor (KNN) for the mortality prediction. The model achieved accuracy of 93%.[9]. Rahul Kumaret al. have proposed another method for COVID-19 prediction through Chest X-Ray image using deep feature learning model. This is a machine learning based classifier using Res Net 152.[10]. In another paper authors have proposed a model for COVID-19 prediction using machine learning and cloud computing. It is a mathematical model based on machine learning to predict threat of COVID-19 in different countries.[11]. R. Sujath at el. have proposed another model for forecasting COVID-19 pandemic in India. They have used linear regression, Multilayer perceptron and Vector auto regression method on kaggle dataset[12]. Another machine learning based model has been proposed to fight against COVID-19. The model can indentifymost susceptible people based on personalized genetic and physiological characteristics.[13] Another very effective model has been proposed to predict the need for ventilation in COVID-19 patients using machine learning framework.[14]. Mohammadreza Nemat et al. have proposed a model which predicts discharge time of COVID-19 patient with survival analysis[15]. Another model for prediction of risk of COVID-19 has been proposed using three machine learning algorithms Artificial Neural Network, Random Forest and Regression Tree [16].

3. Proposed method:

The proposed method is based on machine learning framework. Logistic Regression algorithm has been used for predicting the COVID-19 in someone’s body.

![Flowchart](image-url)

Fig-2: System architecture.
Fig-2 is describing the system architecture.

The symptoms we have considered as: pneumonia, diabetes, chronic obstructive pulmonary disease, asthma, hypertension, cardiovascular disease, renal disease, obesity, tobacco taking habit, and contact with other covid-19 positive one

Step 1: Collection of Data Set: COVID-19 is currently one of the most focused areas of research area. Different web resources are there with plenty of data set. Kaggle is basically the world’s largest data science community from where the COVID-19 patient data has been collected and analyzed for our model[2].

Step 2: Data Processing: The most important part of machine learning based application is processing of data. To prepare an accurate machine learning application data must be properly processed before feeding it to the system. Python machine learning framework has been used for this data processing phase in the proposed model. Followings are the python modules used for our application:

```python
import pandas as pd
import numpy as np
import statsmodels.api as sm
import scipy.stats as st
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.metrics import confusion_matrix
import matplotlib.mlab as mlab
```

For processing the data, it has to be loaded first:

```python
d=pd.read_csv("d:\Research\COVID-19\cvd.csv")
```
Next important phase is segregating the dependent variable from the independent variable as:

```python
import sklearn

new_features=d[['pneumonia','diabetes','copd','asthma','hypertension','obesity','contact_other_covid','covid_res']]
x=new_features.iloc[:,:-1]
```

The last column is the dependent variable, which is identified by -1 index.
Next step for the data processing is the dividing the dataset in training set and testing set. In our model 80% data has been used for training and remaining 20% has been used for testing the model.

```python
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=.20, random_state=5)
```

The fit the data set.

Step 3: Training the System: As described in the above paragraph the data set the segregated in training and testing set, then the model will be trained.

```python
logreg.fit(x_train, y_train)
```

Step 4: Applied Logistic Regression Model: Logistic regression is one of the machine learning algorithm that can be used for classification problem. Proposed system will predict COVID-19 based on Logistic regression.

```python
from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression()
logreg.fit(x_train, y_train)

y_pred = logreg.predict(x_test)
```

4. **Result Analysis:**

The proposed model has given the 92% accuracy based on the dataset, the system is trained.

```python
sklearn.metrics.accuracy_score(y_test, y_pred)
```

Confusion matrix has been computed for analyzing the system performance.
Figure 4: Computation of Confusion Matrix of our proposed model.

```python
In [16]: sklearn.metrics.accuracy_score(y_test, y_pred)
```

```python
Out[16]: 0.921145996868203
```

```python
In [9]: from sklearn.metrics import confusion_matrix
   cm = confusion_matrix(y_test, y_pred)
   conf_matrix = pd.DataFrame(data=cm, columns=['Predicted:0', 'Predicted:1'], index=['Actual:0', 'Actual:1'])
   plt.figure(figsize = (4,4))
   sn.heatmap(conf_matrix, annot=True, fmt='d', cmap='YlGnBu')
```

Figure 5: Confusion Matrix
5. Conclusion:

In this paper a logistic regression based model is being proposed which has considered pneumonia, diabetes, chronic obstructive pulmonary disease, asthma, hypertension, cardiovascular disease, renal disease, obesity, tobacco taking habit, and contact with other covid-19 positive one as the independent variable for covid-19 classification achieving 92% accuracy. In future we have planned for using deep learning mechanism for COVID-19 prediction.

6. References:

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[2] https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset?select=covid.csv
[3] http://www.emro.who.int/health-topics/corona-virus/about-covid-19.html
[4] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200824-weekly-epi-update.pdf?sfvrsn=806986d1_2
[5] P.F. Verhulst, Recherchesmathématiques sur la lois accroissement de la population, M’em. Acad. R. Bruxelles 18 (1845), 3-39.
[6] https://www.google.com/imgres?imgurl=https%3A%2F%2Fstatic.javatpoint.com%2Ftutorial%2Fmachine-learning%2Fimages%2Flogistic-regression-in-machine-learning.png&imgrefurl=https%3A%2F%2Fwww.javatpoint.com%2Flogistic-regression-in-machine-learning&tbmid=LuaHnfur76i8eM&vet=12ahUKEwj15cyjtfbrAhU0j0sFHVTBPVwQMygMegUoARDmAQ..i&docid=maklIDmu8naWM&w=500&h=300&q=logistic%20regression&ved=2ahUKEwj15cyjtfbrAhU0j0sFHVTBPVwQMygMegUoARDmAQ

[7] Hamed A, Sobhy A, Nassar H. Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm. Research Square, 2020. DOI: 10.21203/rs.3.rs-27186/v1.

[8] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jin, Mingyang Zhang, Xiang Huang, Ying Xiao, Haosen Cao, Yanyan Chen, Tongxin Ren, Fang Wang, Yaru Xiao, Sufang Huang, Xi Tan, Niannian Huang, Bo Jiao, Yong Zhang, Ailin Luo, Laurent Mombaerts, Junyang Jin, Zhiguo Cao, Shusheng Li, Hui Xu, Ye Yuan
A machine learning-based model for survival prediction in patients with severe COVID-19 infection
medRxiv 2020.02.27.20028027; doi: https://doi.org/10.1101/2020.02.27.20028027

[9] Mohammad Pourhomayoun, Mahdi Shakibi Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making medRxiv 2020.03.30.20047308; doi: https://doi.org/10.1101/2020.03.30.20047308

[10] Rahul Kumar, Ridhi Arora, Vipul Bansal, Vinod J Sahayashri, Himanshu Buckchash, Javed Imran, Narayan Narayanan, Ganesh N Pandian, Balasubramanian Raman "Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers" medRxiv 2020.04.17.20063461; doi: https://doi.org/10.1101/2020.04.17.20063461

[11] Shreshth Tulia, Shikhar Tulia, Rakesh Tulia, Sukhpal Singh Gill Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing
Internet of Things Volume 11, September 2020, 100222 https://doi.org/10.1016/j.iot.2020.100222

[12] Sujath, R., Chatterjee, J.M. &Hassanien, A.E. A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 34, 959–972 (2020). https://doi.org/10.1007/s00477-020-01827-8

[13] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe and Xi Cheng Artificial intelligence and machine learning to fight COVID-1 https://doi.org/10.1152/physiolgenomics.00029.2020

[14] Hoyt BurdickabCarson Lamec Samson Matarasoac Anna SiefkascGregory BradendR. Phillip Dellingere Andrea McCoy Jean-Louis Vincenttg Abigail Greenc Gina Barnesc Jacob Calvertc Emily PellegrinircRitankar Dasc Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial https://doi.org/10.1016/j.combiomed.2020.103949

[15] Mohammadreza Nemati, Jamal Ansary, Nazafarin Nemati Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data https://doi.org/10.1016/j.patter.2020.100074

[16] Assaf, D., Gutman, Y., Neuman, Y. et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med (2020). https://doi.org/10.1007/s11739-020-02475-0