Prediction of Covid-19 cases using Machine Learning

Sounak Datta[1], Arka Sarkar[2], Iman Saha[3], Subir Baidya[4], Dr. Dharmpal Singh[5], Dr. Sudipta Sahana[6]
[1,2,3,4] *B. Tech, Department of Computer Science & Engineering. JIS College of Engineering, Kalyani, West Bengal, India*
[5] *HOD, Department of Computer Science & Engineering. JIS College of Engineering, Kalyani, West Bengal, India*
[6]*Assistant Professor, Department of Computer Science & Engineering. JIS College of Engineering, Kalyani, West Bengal, India*

Abstract **- Nowadays everyone is being effected from covid-19, In a country with such a huge population like India it is not very easy to test every individual due to shortage of medical kit availability and everyone doesn't have sufficient amount of money and resource. In this paper an effort has been made to design a simple system of detecting Covid-19 based on the symptoms using machine learning. Traditional and ensembled machine learning classifiers have been used, Logistic regression and Decision Tree Classifier is one of them. Logistic Regression showed better results than other ML algorithms by having nearly 97% testing accuracy. A Website was designed by which the users can select the symptoms (Y/N) very easily and also get the result depending on their inputs within a second. The framework made by us can be used, to prioritize testing for covid-19 when testing resources are not sufficient. This system can be used for early detection of covid-19 cases so that the person does not have to go through sever symptoms like difficulty in breathing and does not spread the virus to others.**

Index Terms **- COVID-19, SARS-CoV-2, Machine Learning, Data Analysis, Data Pre-processing, Feature Engineering.**

1.INTRODUCTION

COVID-19 has affected more than 200 countries and has been declared as pandemic by WHO (World Health Organization) in a matter of no time. The virus is transmitted via the respiratory tract when a healthy person comes in contact with the infected person. The infected person shows symptoms within 2–14 days coming in contact to the virus. Currently, some vaccines are developed for preventing this deadly disease, but it would take nearly 1-2 year to vaccinate the total population of India till then we have to take some precautions to prevent this disease. Till April 2021, almost 145 million confirmed cases of coronavirus are detected around the globe. Almost 3 million persons have died, and 125 million persons have recovered from this deadly virus. Since millions and millions are being tested positive everyday around the globe, it is not possible to test all the persons who show symptoms. Machine learning is being used for the identification of novel coronavirus. Machine learning requires a huge amount of data for classifying or predicting diseases and we have collected it online. logistic Regression and Decision Tree Classifier Models for prediction is used. Flask is used for taking the inputs from the users and output the predicted result. If it is detected at early stage and right medical treatment is given at time, then reduction in the death rate around the world is possible.

2.LITERATURE SURVEY

2.1 Machine learning-based prediction of COVID-19 diagnosis based on symptoms.
Author- YazeedZoabi, ShiraDeri-Rozov& Noam Shomron.
They made a prediction models using a gradient-boosting machine model built with decision-tree base-learnering that combine several features to estimate the risk of infection. They established and made a machine-learning approach where the training-validation set consisted of 51,831 tested individuals (of whom 4769 were confirmed to have COVID-19). Their models prediction of COVID-19 test results has high accuracy using only eight binary features. Their model predicted that 0.90 auROC (area under the

receiver operating characteristic curve) with 95% CI: 0.892–0.905. [1]

2.2 Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review
Author- Jiancheng Ye, IoannisApostolopoulos, Aditya Singh Pawar, and ZiyouRenHafsaBareenSyeda, Mahanazuddin Syed, Kevin Wayne Sexton, Shorabuddin Syed, Salma Begum, Farhanuddin Syed, Fred Prior, and Feliciano Yu Jr.
They have made a systematic search of Web of Science, PubMed, and CINAHL databases according to PRISMA rules to identify all potentially relevant studies which have published. They have divided the publication into 3 subjects which are based on AI, and they are Early Detection, Computational Epidemiology, Disease Progression, and diagnosis. The AI techniques they implemented will continue to be used for the monitoring, detection, and containment of the COVID-19 pandemic [5695131]. They have made a recent case study in such a way and chosen to use the Artificial Intelligence methods in the area which is related to containing, tracking, and treating viral infection. Our study provides insights on the prospects of AI on the 3 identified COVID-19 themes—DP, ED, CEE—highlighting the important variables, data types, and available COVID-19 resources. [2]

2.3 Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic
Author- Samuel Lalmuanawma, Jamal Hussain, and LalrinfelaChhakchhuakb
They have used a unique selection of data fir the research article was deployed on the datasets related to the application of Machine learning and AI technology which they used to learn to battle this pandemic. They also have showed up complete reviews of studies on the specified model and on the technology, they are using to fight against the Covid-19 epidemic. The studies also state the types of AI and ML. Therefore it's suggested to return up with a hybrid classification method applying more potential algorithm on multi-database or hybrid-database consisting of clinical, mammographic, and demographic data, as each type of data features a significant factor that would represent truth identity of the infected patients and deployment of the appliance within the world .the use of recent technology with AI and ML dramatically improves the screening, prediction, contact tracing, forecasting, and drug/vaccine development with extreme reliability.[3]

2.4 COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms
Author-Narinder Singh Punn, Sanjay Kumar Sonbhadra, Sonali Agarwal
They have collected everyday generality data of Corona virus 2019 from 22nd January 2020 to 1st April of 2020, from Johns Hopkins University. The dataset which they have collected consists of daily reports of cases. In this case study, they have taken time-series tables in CSV format which have three tables 1. Confirmed 2. Death 3. recovered cases with six attributes and they are state, country/region, last update, recovered cases, confirmed, and death. They have used regression process for the analysis of the pandemic. They have used standard DNN method, and the output layer is made of a single neuron like the RNN. They also made 10% dropout to avoid over fitting problem and final output layer with a single neuron. Thereafter the model which they made on machine learning and deep learning has given a output of the possible number of cases for the next 10 days across the world.[4]

2.5 Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing.
Author- ShreshthTuli, ShikharTuli, RakeshTuli, Sukhpal Singh Gill
They applied an improved mathematical model to analyze and predict the growth of the epidemic. An ML based improved model has been applied by them to predict the potential threat of COVID-19 in countries worldwide. They also deployed the model on a cloud computing platform for more accurate and real-time prediction of the growth behavior of the epidemic. All data up till 4 May 2020 has been used to generate the prediction results in their model. They also presented a case study which shows the severity of the spread of CoV-2 in countries worldwide. Using the proposed Robust Weibull model based on iterative weighting, they wanted to show that their model is able to make statistically better predictions than the baseline Gaussian model. According to their study the baseline Gaussian model is over optimistic.[5]

2.6 Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers

Author- Rahul Kumar, Ridhi Arora, Vipul Bansal, Javed Imran, Vinodh J Sahayasheela, Himanshu Buckchash, Narayanan Narayanan, Ganesh N Pandian, Balasubramanian Raman

Their paper proposes machine learning-based classification of the extracted deep feature using ResNet152 with covid-19 and Pneumonia patients on chest X-ray images. They used SMOTE for balancing the imbalanced data points of covid-19 and Normal patients. Their model got an accuracy of 97.3% on Random Forest and 97.7%using XGBoost predictive classifiers. Such an approach was useful to predict the outbreak at an early stage, which can be used to control the virus. The accuracy on other models were Logistic Regression 96.6%, k-Nearest Neighbour 94.7%, Decision Tree 93.1%, Adaboost 92.1% and Naive Bayes 88.9%.[6]

2.7 Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms

Author-KollaBhanu Prakash, S. SagarImambi, Mohammed Ismail, T Pavan Kumar, YVR Naga Pawan

They have analysed covid-19 datasets to understand which age group is mostly effected due to covid-19. Different types of prediction models were built using machine learning algorithms and their performances were computed and evaluated. Random Forest Regressor and Random Forest Classifier was the best among the other machine learning algorithms in terms of accuracy. They have categorized symptoms in three category most common, moderate, severe. They have analysed that the age groups of 20-50 are highly probable of getting infected with COVID-19 nearly 63%.[7]

3. RESEARCH METHODOLOGY

In this paper shows the analysis of various machine learning algorithms, the algorithms used in this paper are Logistic Regression, Decision Tree Classifier Models, Naive Bias Classifiers which can be helpful for medical analysts for accurately diagnose of Covid19. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology Covid-19 Predictor consists of 6 steps.

1. data collection is being performed and
2. data analysis
3. Understanding the data
4. data preprocessing
5. Feature Engineering
6. Analysis of the Preprocessed Data

Machine Learning

There are several supervised machine learning models that measure accustomed Classify text into this half. The machine learning algorithms like support vector machine (SVM), multinomial Naive mathematician (MNB), supply regression, call tree were used for acting this task.

Decision Tree

Decision Tree algo is within the sort of a multidimensional Chart wherever the inner node represents the dataset attributes and also the outer branches treated as the end result. Decision Tree is chosen as a result of they're quick, reliable, simple to interpret and really very little information preparation is needed. In decision Tree, the prediction of category label originates from root of the tree. The figure of the basis attribute is compared to records attribute. On the results of comparison, the $P(A|B) = (P(B|A)P(A)) / P(B)$ (1)

Logistic regression

This rule predicts the category of numerical variable supported its relationship with the label. The rule typically calculates the category membership likelihood. Logistic Regression largely used for binary classification issues. In logistic regression rather than fitting a line or hyper plane, the logistic regression rule uses the function operate to squeeze the output of an equation between zero and one.

Naive Bayes

Naive Bayes rule relies on the Bayes rule. The independence between the attributes of the dataset is that the main assumption and also the most vital in creating a classification. it's simple and quick to predict and holds best once the idea of independence holds. The theory calculates the posterior chances of an incident (A) given some previous l of even chances B drawn by P(A/B).

4.IMPLEMENTATION

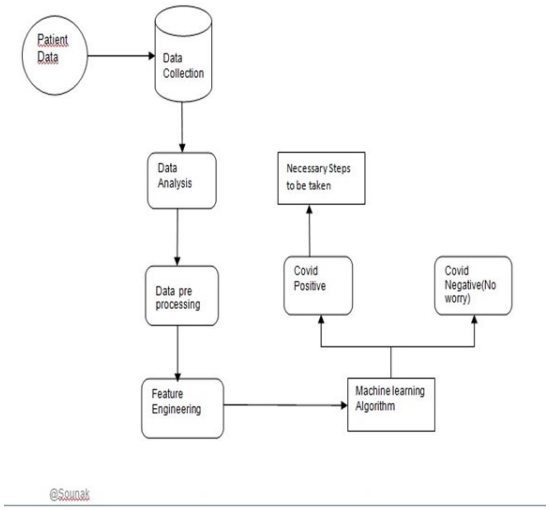

Fig1. Implementation process

Symptoms:

1. Cough (Yes/No).
2. Fever (Yes/No).
3. Sore throat (Yes/No).
4. Running Nose (Yes/No).
5. Headache (Yes/No).
6. Asthma (Yes/No).
7. Chronic Lung Disease (Yes/No).
8. Heart Disease (Yes/No).
9. Diabetes (Yes/No).
10. Hypertension (Yes/No).
11. Fatigue (Yes/No).
12. Gastrointestinal (Yes/No).
13. Contact with COVID Patient during last 14 days (Yes/No).
14. Attended Large Gathering (Yes/No).

```
[3]: df.head()
```

[3]:

	Breathing Problem	Fever	Dry Cough	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	...	Fatigue	Gastrointestinal	Abroad travel	Contact with COVID Patient
0	1	1	1	1	1	0	0	0	0	1	...	1	1	0	1
1	1	1	1	1	0	1	1	1	0	0	...	1	0	0	0
2	1	1	1	1	1	1	1	1	0	1	...	1	1	1	0
3	1	1	1	0	0	1	0	0	1	1	...	0	0	1	0
4	1	1	1	1	1	0	1	1	1	1	...	0	1	0	1

5 rows × 21 columns

Fig 2: Dataset after implementing feature engineering

```
corr=df.corr()
sns.heatmap(corr)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x16aec486198>
```

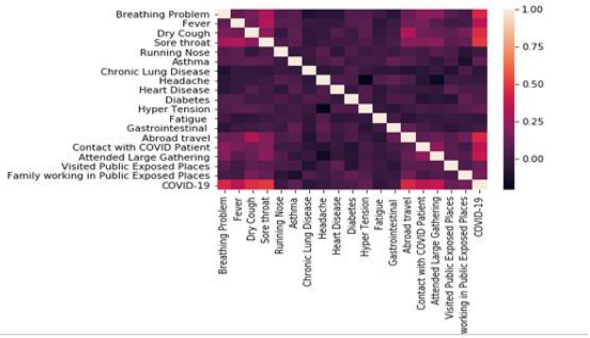

Fig 3: Analysis of Preprocessed Data

```
sns.countplot(df['COVID-19'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x16aea740dd8>
```

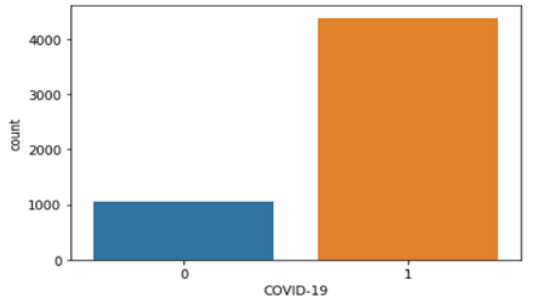

Fig 4: Co-relation between the features and the target value

5.RESULT AND ANALYSIS

Algorithm	Precision	Recall	F1 score	Accuracy (%)
Logistic Regression	0.97	0.98	0.98	97.05
Multinomial Naive Bayes	1.0	0.71	0.82	77.36

Table 1: Analysis of traditional machine learning algorithms

Fig 5: Comparative Analysis of machine learning models

A windows system with 8 GB Ram and 2.0 GHz processors is used for performing this particular work. Scikit-learn tool has been used for performing machine learning classification with the help of various libraries. After performing data analysis deeper insights about the data were achieved. The data is being split into 80:20 ratio where 80% data is being used for training the model and 20% is used for testing the model. The classification models were created by supplying the features which were extracted from the feature engineering step. In order to achieve more accuracy, explore the generalization of our model from training data to predicted data and reduce the possibility of overfitting, the original dataset have been split into separate training and testing sub datasets. Cross-validation strategy has been conducted for all algorithms and this process was repeated many times independently to avoid the problems like overfitting or selection bias. The results showed that logistic regression shows better result than all other algorithms by having 97% accuracy. Multiple regression and classification models have been applied and being observed the differences among them. After that the best possible model for the prediction has been taken.

6.CONCLUSION

There is no particular drug for curing Covid-19 and there is also a shortage for the vaccines. So in this type of situations with the increasing number of deaths due to this pandemic, it has become mandatory to develop a system to predict Covid-19 effectively and accurately. The motivation for the paper was to find the most efficient ML algorithm for detection of Covid-19 and implement that into a website for the users. This study compares the accuracy scores of Logistic Regression and Naive-Bayes classifier. The result of this study indicates that the Logistic Regression is the most efficient algorithm with accuracy score of 97% for prediction of Covid-19. The efficiency of the models can be improved by increasing the quantity of patient's data.

REFERENCE

[1] YazeedZoabi, ShiraDeri-Rozov, Noam Shomron (2021),Machine learning-based prediction of COVID-19 diagnosis based on symptoms: https://www.nature.com/articles/s41746-020-00372-6

[2] Jiancheng Ye, IoannisApostolopoulos, Aditya Singh Pawar, and ZiyouRen HafsaBareenSyeda, Mahanazuddin Syed, Kevin Wayne Sexton, Shorabuddin Syed, Salma Begum, Farhanuddin Syed, Fred Prior, and Feliciano Yu Jr (2021), Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7806275/

[3] Samuel Lalmuanawma, Jamal Hussain, and LalrinfelaChhakchhuakb (2020), Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7315944/

[4] Narinder Singh Punn, Sanjay Kumar Sonbhadra, Sonali Agarwal (2020), COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms: https://www.medrxiv.org/content/10.1101/2020.04.08.20057679v2.full

[5] ShreshthTuli, ShikharTuli, RakeshTuli, Sukhpal Singh Gill (2020), Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing: https://www.sciencedirect.com/science/article/abs/pii/S254266052030055X

[6] Rahul Kumar, Ridhi Arora, Vipul Bansal, Javed Imran, Vinodh J Sahayasheela, Himanshu Buckchash, Narayanan Narayanan, Ganesh N Pandian, Balasubramanian Raman (2020), Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers: https://www.medrxiv.org/content/10.1101/2020.04.13.20063461v1.full

[7] KollaBhanu Prakash, S. SagarImambi, Mohammed Ismail, T Pavan Kumar, YVR Naga Pawan (2020), Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms: Ananthi2021_J._Phys__Conf. _Ser._1767_012006.pdf

[8] Cheng Jin, Weixiang Chen, Yukun Cao, Zhanwei Xu, Zimeng Tan, Xin Zhang, Lei Deng, Chuansheng Zheng, Jie Zhou, Heshui Shi, Jianjiang Feng (2020), Development and Evaluation of an AI System for COVID-19

Diagnosis: https://doi.org/10.1101/2020.03.20.20039834

[9] Song Ying,Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Huiying Zhao, Ruixuan Wang, Yutian Chong, ProfileJun Shen, Yunfei Zha, Yuedong Yang (2020), Deep learning Enables Accurate Diagnosis of Novel Coronavirus : https://doi.org/10.1101/2020.02.23.20026930

[10] Ensheng Dong, Hongru Du, Lauren Gardner (2020), An interactive web-based dashboard to track COVID-19 in real time:https://doi.org/10.1016/S1473-3099(20)30120-1

[11] Covid-19 Symptom Data: https://www.kaggle.com/iamhungundji/covid19-symptoms-checker (2020)

[12] Description of Logistic Regression Algorithm (2019) https://machinelearningmastery. com/logistic-regression-for-machine-learning

[13] AnhvinhDoanvo, XiaoluQian, DivyaRamjee, Helen Piontkivska, Angel Desai, Maimuna Majumder (2020), Machine Learning Maps Research Needs in COVID-19 Literature: https://www.sciencedirect.com/science/article/pii/S2666389920301641