Implementation of Machine Learning Models to Differentiate the Symptoms Of COVID-19 and Risk Stratification of Disease Severity

Amosh Sapkota[1], Srikanth Viswanadhuni[2], Anand Kumar[3] , Kalyan Shankar[4] , Anjali Mathur[5] .
[1,2,3,4]Students, Department of Computer Science and Engineering, Koneru Lakshmaiah
Education Foundation, Guntur (A.P), India, 522502.
[5]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah
Education Foundation, Guntur(A.P), India , 522502**.**

***Abstract* : -In the current situation due to the similar symptoms of both covid-19 and flu many people are unaware between covid-19 and flu which may lead to demise of a person. So sort of methods are required to classify the symptoms between covid-19 and other disease to control the demise rate. Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. In the lighting view of current pandemic situation diagnosis of these disease is done only through some clinical tests like RT-PCR, CT- Scan of lung images to identify the covid-19 since these tests take both much time and also very expensive, we will be implementing a solution to overcome these current problems faced by people in the pandemic situation. After taking a literature survey we got that image processing , data mining, machine learning, pattern classification are highly used methods to get solution of these problem.**

Key words : Covid-19, flu, respiratory illness, cardiovascular disease, diabetes, chronic respiratory disease, RT-PCR, CT-Scan, image processing, data mining, machine learning, pattern classification.

1. INTRODUCTION :

1.1 Novel Corona Virus disease (COVID-19) :

From 2019, the Corona Virus showed severe influence and became a cause of many casualties and a large number of deaths. The world level pandemic was declared with the name of COVID-19. Initially, COVID-19 was showing symptoms like flu, which then turned to pneumonia and finally infected the lungs. But the cause of demise was a heart attack for the majority of COVID-19 patients. The person affected by the virus started showing symptoms within 2 to 14 days. In short, the COVID-19 showed mixed symptoms of multiple diseases. Sometimes, COVID positive patients were shown full recovery, and started living a healthy life but in most cases, the situation turns severe and patients are kept in ICU. There was a high requirement to find out the recovery rate and severeness of COVID-19.

1.2 Machine learning techniques :

In the proposed work, we try to differentiate COVID-19 symptoms other than flu by using various Machine Learning models and after getting those symptoms, we include them in clinical symptoms of COVID-19 and classify the dataset into three groups. They are classified as Mild, Moderate, and Severe Patients respectively. As COVID-19 showed symptoms of various diseases, the dataset grows bigger. Machine Learning algorithms are a great choice for a bigger dataset. We used K-Nearest Neighbors, Logistic Regression and Decision tree algorithms for differentiating COVID-19 with flu. Decision tree, Naïve Bayes and K-Means algorithms are used for risk stratification.

2. LITERATURE SURVEY :

Research work was carried out [1]with the methods such as Real-time data query was carried out, visualized in their website ant then the queried data is used for Susceptible-Exposed Infectious-Recovered (SEIR) predictive modelling. The author utilized SEIR modelling to forecast the COVID-19 outbreak within and outside of China based on daily observations. The author also analysed the queried news and classified the news into negative and positive sentiments, to understand the influence of the news on people’s behaviour both politically and economically. According to his findings, the news queried in their system, they found that there are more negative articles than positive articles and displayed similar words for both negative and positive sentiments. The top five positive articles are about collaboration and strength of individuals in facing this epidemic, and the top five negative articles are related to uncertainty and poor outcome of the disease such as deaths. Finally, it was concluded that there is still an unclear infectious disease, which means the author can only obtain an accurate SEIR prediction after the outbreak ends. Another research work discusses[2] about the mathematical model of the spread of population structured by age. As the disease spreads through social contact and varies on age, it is important to predict the spread of disease by the change in the social structure. The mathematical model was a combination of contact structure compilation along with empirical case data which is used to assess the impact of the COVID-19 pandemic. The model shows that a sustained period of lockdowns with periodic relaxation will reduce the number of cases. Mathematical model contains both symptomatic and asymptomatic infectives. Certain research work was carried out[3] regarding novel Coronavirus (COVID-19) in India. In this paper, the author delved into the origin of COVID-19 and classified the disease into different categories. Coming to the origin of the disease, the Novel Corona Virus earlier known only as the Wuhan virus, expanded its circle in South Korea, Japan, Italy, Iran, and finally spreading its routes to India. It was given the name novel as it was never seen before mutation of an animal Coronavirus. The disease COVID-19 is very similar in symptomatology to other viral respiratory infections. Cases vary from mild forms to severe ones that can lead to serious medical conditions or even death. The incubation period for the virus to be present in the human body is 2-14 days 3 but it is still unknown. There are some clinical syndromes associated with COVID-19 infection which are divided into Mild and Severe pneumonia. Diagnosis of the COVID-19 disease can be done in two ways. They are RT-PCR test and CT-SCAN of lungs. The results of CT-SCAN are divided into different categories such as Mild, Moderate and Severe ARDS. Sepsis and Sepsis shock are the last two stages of results which pose great difficulty to identify the disease. A common precaution to prevent the COVID-19 virus is to clean your hands. The research work explains[4] about the machine learning algorithms which were used for classifying clinical reports into four different classes namely COVID, ARDS, SARS and BOTH (SARS and ARDS). The algorithms used were Logistic regression and Multinomial Naïve Bayes. Dataset was used for text analysing which is changed in the form of Natural Language Processing. Clininal notes and findings were used for data related to text mining. After the classification, it was noticed that the multinomial Naïve Bayesian model shows the best results by having 94% precision. In this research paper[5], deep learning models were proposed for predicting the number of COVID-19 positive cases in Indian states. Data analysis and an increase in the number of cases in India was observed. States are categorised based on daily growth, the number of cases into severe, moderate and mild. Recurrent neural Network (RNN) based long short-term memory (LSTM) cells were used for prediction. LSTM cases were tested on various states and based on absolute error, the model which shows maximum accuracy is considered. Bi-directional LSTM gives the best accuracy while convolutional LSTM gives the worst results. This paper shows[6] the prediction of COVID-19, which was done, due to the high level of uncertainty and lack of crucial data, standard models have shown low accuracy for long-term prediction. The dataset was used from the WHO website and was dated till January 2020. The modelling strategy is formed around the assumption of transmitting the infectious disease through contacts and considering three different classes of well-mixed populations; susceptible to infection (class S), infected (class I), and the removed population (class R is shown to those who have recovered). It is further assumed that the class I transmits the infection to class S where the number of probable transmissions is proportional to the total number of contacts. The number of individuals in class S progresses as a time series, often computed using a basic differential equation. 4 In this paper[7], the present analysis was carried out based on the publicly available data of the newly confirmed daily cases reported from 11th of January till the 10th of February. The method used was the estimation of the mean values of the main epidemiological parameters, that is the basic reproduction number (Ro), the case fatality, and case recovery ratios, along with their 90% confidence intervals. In the second scenario, results are derived by taking twenty times the number of reported cases for the infected and forty times the number for the recovered cases, while keeping constant the number of deaths. The basic reproduction number (Ro) is one of the key values that can predict whether infectious disease will spread into a population or die out. The paper discusses[8] about the size of epidemic in Wuhan which was estimated based on the number of cases exported from Wuhan to cities outside China and forecast of the domestic and global public health risks of epidemics was carried out. A metapopulation susceptible exposed infected recovery (SEIR) model was used to simulate epidemics across all major cities in China. Markov Chain Monte Carlo method was used to estimate basic reproduction number and presented using the resulting posterior mean and 95% credible interval. Author first inferred the basic reproductive number of COVID-19 and the outbreak size in Wuhan from Dec 1st, 2019 to Jan 25th, 2020 based on the number of cases exported from Wuhan to cities outside mainland China then the number of cases that had been exported from Wuhan to other cities in mainland China. In this way the spread of COVID-19 within and outside China was forecasted. By the time research was conducted COVID-19 might have already spread to other major cities inside and outside China and made new hotspot there. Another research work[9] focuses on a classification model using deep feature extraction and Support vector machine. From the dataset containing X-ray images of COVID-19 patient, pneumonia patient and healthy people; X-ray images of COVID-19, pneumonia patient and healthy people were classified. First the author screened out healthy people and pneumonia patients. As mostly pneumonia symptoms and COVID-19 are somehow similar, and both are lung related diseases; pneumonia patients were screened out from COVOD-19 patients. Pretrained networks such as AlexNet, XceptionNet, ShuffleNet, GoogleNet, VGG16, VGG19, ResNet18, ResNet50, ResNet101, InceptionV3 etc. were used for deep feature extraction. The deep features obtained from these networks were fed to the SVM for classification. Then, the traditional image classification methods were applied for the detection of COVID-19. The 5 proposed study used pre trained CNN models for the detection of COVID-19. Only X-Ray image processing can't rule out COVID-19.

3. METHODOLOGY :

3.1 K-Nearest Neighbor Classifier :

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning techniques. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using the K- NN algorithm.

Mathematical formulas for K- Nearest Neighbor classification is mentioned below.

Distance function of Euclidean

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Distance function of Manhattan

$$d(x,y) = \sum_{i=1}^{m}|x_i - y_i|$$

Distance function of Minkowski

$$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p}$$

3.2 Decision Tree Classifier :

Decision Tree is a machine learning model which can be used for both classification and Regression problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.

Decision nodes are used to make any decision and have many branches, whereas Leaf nodes are the output of those decisions .

The mathematical formulas of decision tree classifier are given as follows

1. Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)
2. Gini Index= 1- $\sum_j P_j^2$

3.3 Logistic Regression :

Logistic regression is one of the supervised machine learning model algorithms which is used to predict the probability of a target variable. Here the nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In the regression model there are two types of variables -dependent, independent. The dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X

Logistic regression formula

P(X) = P(Y=1|X)

Logistic regression equation

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x))

3.4 Naïve Bayes :

Naïve bayes algorithm works on Bayes theorem which is one of the classification technique of machine learning model . A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The Naive Bayes model is easy to build and particularly useful for very large data sets.

Mathematical formula of Naïve Bayes model is mentioned below

P(c|x) = p(x|c) p(c) / p(x)

3.5 K- Means :

K means is an iterative algorithm that tries to partition the dataset into *K* pre-defined clusters where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different as possible. It works on the principle of calculating the sum of the square of distances between the two data points and the cluster's centroid is at a minimum.

The objective function is:

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{c} (|Xi - Vj|)\hat{}2$$

‘$||x_i - v_j||$’ is the Euclidean distance between x_i and v_j
‘c_i’ is the number of data points in i^{th} cluster.
‘c’ is the number of cluster centres.

4 . RESULTS AND ANALYSIS :

Here in these paper we have conducted experiments on some of the techniques to predict covid-19 and stratification of severity of covid-19 disease.

S. No	Name of machine learning algorithm	Accuracy
1	K- Nearest Neighbor Classifier	0.86
2	Decision Tree Classifier	0.90
3	Logistic Regression	0.89
4	Naïve Bayes	0.94
5	K- Means	0.80

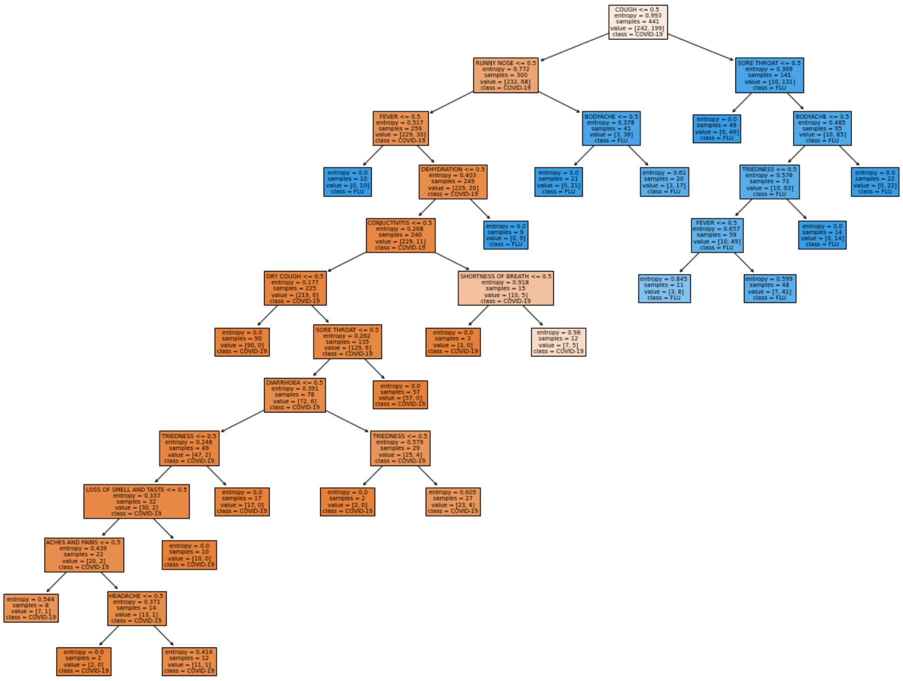
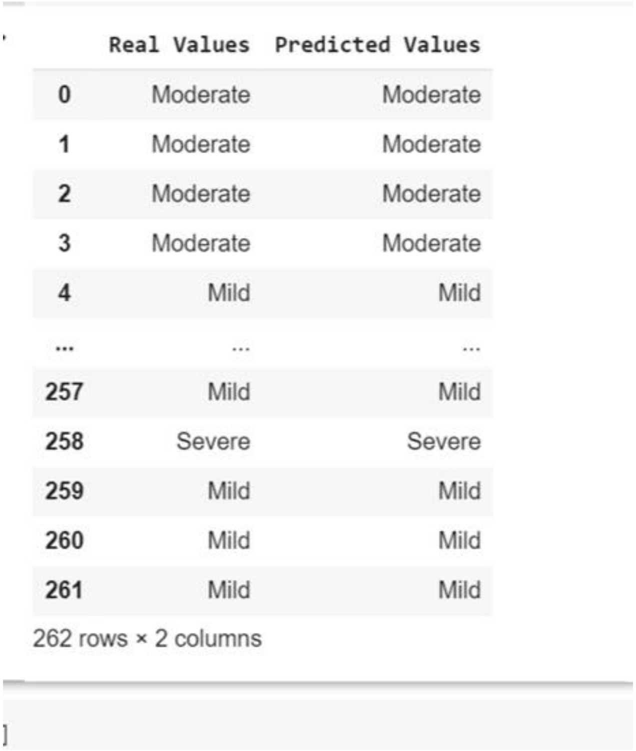

COUGH <= 0.5 entropy = 0.993 samples = 441 value = [242, 199] class = COVID-19
RUNNY NOSE <= 0.5 entropy = 0.772 samples = 300 value = [232, 68] class = COVID-19
SORE THROAT <= 0.5 entropy = 0.369 samples = 141 value = [10, 131] class = FLU
FEVER <= 0.5 entropy = 0.517 samples = 259 value = [229, 30] class = COVID-19
BODYACHE <= 0.5 entropy = 0.378 samples = 41 value = [3, 38] class = FLU
entropy = 0.0 samples = 46 value = [0, 46] class = FLU
BODYACHE <= 0.5 entropy = 0.485 samples = 95 value = [10, 85] class = FLU
entropy = 0.0 samples = 10 value = [0, 10] class = FLU
DEHYDRATION <= 0.5 entropy = 0.403 samples = 249 value = [229, 20] class = COVID-19
entropy = 0.0 samples = 21 value = [0, 21] class = FLU
entropy = 0.61 samples = 20 value = [3, 17] class = FLU
TRIEDNESS <= 0.5 entropy = 0.576 samples = 73 value = [10, 63] class = FLU
entropy = 0.0 samples = 22 value = [0, 22] class = FLU
CONJUCTIVITIS <= 0.5 entropy = 0.268 samples = 240 value = [229, 11] class = COVID-19
entropy = 0.0 samples = 9 value = [0, 9] class = FLU
FEVER <= 0.5 entropy = 0.657 samples = 59 value = [10, 49] class = FLU
entropy = 0.0 samples = 14 value = [0, 14] class = FLU
DRY COUGH <= 0.5 entropy = 0.177 samples = 225 value = [219, 6] class = COVID-19
SHORTNESS OF BREATH <= 0.5 entropy = 0.918 samples = 15 value = [10, 5] class = COVID-19
entropy = 0.845 samples = 11 value = [3, 8] class = FLU
entropy = 0.599 samples = 48 value = [7, 41] class = FLU
entropy = 0.0 samples = 90 value = [90, 0] class = COVID-19
SORE THROAT <= 0.5 entropy = 0.262 samples = 135 value = [129, 6] class = COVID-19
entropy = 0.0 samples = 3 value = [3, 0] class = COVID-19
entropy = 0.98 samples = 12 value = [7, 5] class = COVID-19
DIARRHOEA <= 0.5 entropy = 0.391 samples = 78 value = [72, 6] class = COVID-19
entropy = 0.0 samples = 57 value = [57, 0] class = COVID-19
TRIEDNESS <= 0.5 entropy = 0.246 samples = 49 value = [47, 2] class = COVID-19
TRIEDNESS <= 0.5 entropy = 0.579 samples = 29 value = [25, 4] class = COVID-19
LOSS OF SMELL AND TASTE <= 0.5 entropy = 0.337 samples = 32 value = [30, 2] class = COVID-19
entropy = 0.0 samples = 17 value = [17, 0] class = COVID-19
entropy = 0.0 samples = 2 value = [2, 0] class = COVID-19
entropy = 0.605 samples = 27 value = [23, 4] class = COVID-19
ACHES AND PAINS <= 0.5 entropy = 0.439 samples = 22 value = [20, 2] class = COVID-19
entropy = 0.0 samples = 10 value = [10, 0] class = COVID-19
entropy = 0.544 samples = 8 value = [7, 1] class = COVID-19
HEADACHE <= 0.5 entropy = 0.371 samples = 14 value = [13, 1] class = COVID-19
entropy = 0.0 samples = 2 value = [2, 0] class = COVID-19
entropy = 0.414 samples = 12 value = [11, 1] class = COVID-19

Figure 1 : Visualization of Decision Tree Classifier .



	Real Values	Predicted Values
0	Moderate	Moderate
1	Moderate	Moderate
2	Moderate	Moderate
3	Moderate	Moderate
4	Mild	Mild
...	...	...
257	Mild	Mild
258	Severe	Severe
259	Mild	Mild
260	Mild	Mild
261	Mild	Mild

262 rows × 2 columns

Figure 2 : Screenshot of real values and predicted values

By the end of the implementation of the given machine learning models, we have reached the following conclusions respectively. Starting with the accuracy of the K-NN model is 0.8611. Its precision is 0.86, recall value is 0.90, f1-score is 0.88 and support is 20. Secondly, the accuracy of the Decision tree classifier model is 0.905. Its Precision 0.89, recall is 0.91, f1-score is 0.90 and support is 93. Finally, the accuracy of the Logistic Regression model is 0.89. Its precision is 0.90. Prediction to classify the difference between COVID-19 and flu is a difficult task because it is hard to find out the symptoms of COVID-19. In the case of stratification of the data, by the end of the implementation of the given machine learning models, we have reached the following conclusions respectively. Starting with the accuracy of the Decision tree model is 0.94. Its precision is 0.93, recall value is 0.93, f1-score is 0.92 and support is 87. Secondly, the accuracy of the Naïve Bayes classifier model is 0.94. Finally, the accuracy of the K Means model is 0.80.

5 . CONCLUSION :

From the models accuracies obtained for the dataset through various machine learning algorithms, we can say that the decision tree classifier gives the best results, which is 0.905 followed by Logistic regression and KNN model to the least accurate respectively. When it comes to the stratification of the data, the decision tree and Naïve Bayes are almost the same as giving the best results at the accuracy of 0.94 while K Means model gives the least accurate results. Hence, we conclude that our work on COVID-19 prediction and its severity differentiation has been completed, which is used to reduce the time and cost for the doctors to diagnose the disease and find which stage the patient is at. The limitation of this project is that COVID-19 is an evolving disease showing different types of symptoms at different points in time. This work may not be reliable if the symptoms are completely different from the dataset we used at this moment. Another drawback is that there are really few or no works containing these machine-learning techniques to compare them for their accuracy.

6 . REFERENCES :

[1] Fairoza Amira Binti Hamzaha, Corona Tracker Community Research Group , COVID 19 site of WHO , 2020.

[2] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China

[3] Joseph T Wu, Kathy Leung, Gabriel M Leung, Nowcasting and forecasting the potential domestic and international spread of the nCoV-19 outbreak originating in Wuhan, China: a modelling study, 20

[4] Organization WH. WHO Statement Regarding Cluster of Pneumonia Cases in Wuhan, China; 2020. Available from: https://www.who.int/china/news/detail/ 09-01-2020-whostatement-regarding-cluster-of-pneumonia-cases-in-wuhan-china.

[5] Prabira Kumar Sethy, Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine, 2020

[6] S. Towers and Z. Feng, "Social contact patterns and control strategies for influenza in the elderly"

[7] The Johns Hopkins Center for Health Security. Daily updates on the emerging novel coronavirus from the Johns Hopkins Center for Health Security. February 9, 2020; 2020. Available from: https://hub.jhu.edu/2020/01/23/ coronavirus-outbreak-mapping-tool-649- em1-art1-dtd-health/

[8] Varsha Kachroo , Novel Coronavirus (COVID-19) in India: Current Scenario ,International Journal of Research and Review , vol 7;Issue: 3,March 2020.

[9] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china.