A Prediction of Corona Disease Transmission Using A Traditional Machine Learning Approach

Chandran Venkatesan, Sumithra M G, Karthick A and Anto Merline M

Department of Electronics and Communication Engineering
KPR Institute of Engineering and Technology, India
sumithra.mg@kpriet.ac.in, karthick.power@gmail.com, merlinemanoharan@gmail.com

Elakkiya Balan
Department of Electronics and Communication Engineering
Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India, elakz@gmail.com

Jayarajan V
Associate Consultant
ATOS Syntel, India
jayarajanv25@gmail.com

Abstract - In this current scenario, covid pandemic breaks analysis is becoming popular among the researchers. The various data sources from the different countries analyzed to predict the possibility of coronavirus transition from one person to another person. The datasets are not providing more information about the causes of the corona. Many authors provided the solution by using chest X-ray and CT images to predict the corona. In this paper, the covid pandemic transition process from one person to another person was classified using traditional machine learning algorithms. The input labels are encoded and transformed, utilizing the label encoder technique. The XG boost algorithm was outperformed all the other algorithms with overall accuracy and F1-measure of 99%. The Naïve Bayes algorithm provides 100% accuracy, precision, recall, and F1-Score due to its improved ability to handle lower datasets.

Keywords - Pandemic, Machine Learning (ML), Corona, XG Boost, Naïve Bayes.

1. Introduction
The COVID-19 epidemic was caused by a type of virus called coronavirus starting from December 2019. Here coronaviruses (CoVs) are arranged by a large cluster of viruses. The coronavirus is of many types, and it affects not only humans but also birds and animals. Some kind of coronavirus causes mild respiratory syndrome among humans [1]. Therefore coronavirus is not newly created, but the virus that causes covid-19 is new. The virus that causes coronavirus is named as SARS-CoV-2. It has been originated from bats. This virus has a unique ability to spread from one species to another, including human-to-human transmission. The new type of viruses which were originated from bats was similar to this kind of pandemic diseases known as Severe Acute Respiratory Syndrome (SARS) in 2002, from Guangdong, China and Middle Eastern Respiratory Syndrome (MERS) in 2012, and finally SARS-CoV2-2 in 2019, from Wuhan, China. All these types of viruses can infect people and highly spreadable [2].

The outbreak of the virus has changed the lives of humans globally by affecting every individual physiologically and psychologically. As of 17th May 2020, the World Health Organization (WHO) reported that there had been currently 4,494,873 confirmed cases and 305,976 deaths all over the world. From the coronavirus outbreak, the death rate is continuously increasing as more infections, and as a result, deaths have become unstable to measure [3]. The general signs and symptoms of coronavirus are dry cough, fever, shortness of breath, fatigue, headache, sore throat, and so on. The symptoms are generally present in humans for 2 to 14 days after a person is infected with a virus [4]. Presently, no antiviral treatment for this virus has been medically proven to be made useful.

Moreover, by taking some preventive measures will temporarily stop the spreading of the virus from one person to another person. Some of the preventive measures one should be followed are washing hands frequently, covering mouth while coughing and sneezing, avoiding close contact, avoid visiting public places, and, most importantly keep a distance from unknown odd objects. The lifecycle of humans will cause a drastic change in their working activities instead of their routine habits due to lockdown in various countries by government order. The increase in the number of cases has become severe stress on the medical community, most importantly supplying a lack of intensive care asserts.
To solve the above need, data for coronavirus studies are available from many government organizations, hospitals, and other sources. That includes how the virus spreads and is being treated. There are two main types of data available. One is called textual data. The COVID-19 consists of over 25000 technical articles and the data about adverse effects and so on. Humans can't read through all the literature and extract critical information.

The second type of data is numerical data or semi-numerical data. Data is being added each day to these datasets. A more granular data set at a patient-level includes additional information such as the patient's symptoms, other medical conditions, age, gender, and so on. The essence of the table is that each row becomes a point in multi-dimensional space, and all Machine Learning (ML) algorithms look for patterns in that space. This research work supports predicting the COVID-19 situation by considering the Korean dataset [5]. The prediction helps to project the spread of the virus over time and put more robust medication measures in its paths or estimate which health centers will be overwhelmed so that we can be prepared ahead of time with better processing in supplies for the medical community.

Considering the ML approaches, the system or a computer program will learn or make it adapting the data through experiences. Therefore based on the traditional ML concept, the design of the system is obtained from the model of the given data [6]. Hence this is the output of the machine-learned program. It is the system that generated a piece of code or a model and also an internal representation of data that the model is learned by using ML algorithmic approaches. There are several ML applications over the past years, and hence it is one of the significant areas of research. Some of the ML applications used recently are object recognitions to recognize and tag objects, document classification, anomaly detection, fraud detection, gaming’s, etc. Using ML algorithms, we will able to predict how a particular epidemic is spreading through a neighborhood or a system. The ML algorithms, such as K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), and eXtreme Gradient Boosting (XGBoost) algorithms are applied.

2. Proposed Work

The ML workflow is shown in the below Figure 1. The primary step of the machine learning workflow is to determine the problem statement. The typical steps are clearly explained in the below Figure 1. If the performance metrics are meeting the desired problem formulation, the parameter tuning has to be done.

Traditional ML algorithms

The ML algorithms are classified broadly as supervised, unsupervised, and reinforcement learning. In this work, supervised learning with classification tasks such as Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), K-Nearest Neighbours (KNN) and some ensemble methods like Random Forest (RF) and XG Boosting (XGB) algorithms are applied.

![Fig. 1: Machine learning workflow process](image)

Naive Bayes

The Naive Bayes classifier is an algorithm that works as a simple probabilistic classifier based on Bayes theorem with strong independence assumptions between feature values. It is widely used in text classification in turns provides better accuracy for a limited amount of data. The Bayesian network tree constructed depends on the conditional probability of happening. The Naive Bayes classification is formulated in the below Equation (1).

\[
\text{Naive Bayes}
\]
Where,

\[ y = \text{class label (released and isolated)} \]
\[ x = \text{input features values}. \]
\[ P(y|x) = \text{Posterior probability of class } y \text{ given target values}. \]
\[ P(y) = \text{class probability} \]
\[ P(x|y) = \text{likelihood defines the probability of predicted class} \]
\[ P(x) = \text{prior probability of the predictor}. \]

The pseudo-code for the Naïve Bayes algorithm is shown below in Figure 2.

**Fig. 2: Naïve Bayes algorithm**

**Decision Tree**

A decision tree is a tree-based algorithm used to solve both the classification and regression problems. The Decision tree starts with the root node and passes it into the decision node. In this work, classification and regression tree algorithms are used. The Gini index as the metric for classification, to determine how well the feature data values together mixed in nature. The Gini score determines the mixing of data points. If it results in high value, the data points are unequally combined, and for minimum value data points are equally mixed. The Gini index formula is written as given in Equation (2),

\[ \text{Gini index} = 1 - \sum_{c} P_{c}^{2} \]  

(2)

The below Figure 3 show sample decision tree structure with Gini index value for the COVID-19 classification data.

**Fig. 3: The decision tree structure for deriving root node**

**Support Vector Machine**

Support vector machines are the most popular machine learning algorithms. It is used for both classification and regression as a support vector classifier and support vector regressor, respectively. The major drawback of using the support vector machines is it takes too long to process. The idea behind the SVM is to find the hyperplane that best separates the features into different domains. The support vector machines have various kernel tricks such as linear, polynomial, and radial basis function/ Gaussian kernel. In this work, the linear kernel trick was used due to most of the features being linearly separated, and also, it is very fast in nature.

**Pseudo-code**

The pseudo-code algorithm is shown below in Figure 4.

**Fig. 4: Pseudo-code**

**Fig. 5: Pseudo-code**

The following figures show the region of the curve for the kernel tricks, such as linear, polynomial, and radial basis function curves, respectively. A ROC curve (receiver operating characteristic curve) is used to plot the two parameters True Positive Rate (TPR) and False Positive Rate (FPR) for providing the classification model performance at different thresholds level.

The Figure 6(a), 6(b), 6(c) represents the polynomial kernel trick, linear kernel trick, and radial basis kernel trick, respectively. From this plot, we can calculate the learning curve for different threshold values.
Random Forest(RF)

Random forest algorithm is like a bootstrapping algorithm with a decision tree model. The tree built with different observations. The various trees grown for different instances based on the majority voting the final class predicted. The random forest accuracy will increase by tuning the hyper parameters such as the number of decision trees, maximum depth of the tree. These random forest algorithms are relatively fast and robust algorithms for classification. Here the random forest algorithm has better accuracy. The Figure 7 shows the examples of how the random forest algorithm built with different decision trees and predicting the final class based on the majority voting mechanism. Let us assume the data was fed into the algorithm. The random forest algorithm built three trees the first tree predicting the class as A, tree 2 predicting the class as B and tree 3 predicting the class as B. By a majority of voters, the algorithm classifies the data as class B.

XGBoost algorithm

XG Boost algorithm is one of the methods of decision trees based on the ensemble method type of machine learning algorithm based on the framework of gradient boosting. The XGboost algorithm will outperform when the data is a small structured/tabular format. The objective function will be used to minimize the loss function and regularization. It is given below in Equation (3).

\[ L(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + f_t(X_i) + \Omega(f_t) \]  

3. Experimental Results and Discussion

Dataset Description

COVID-19 dataset was collected from the Korea Centers for Disease Control and Prevention. The dataset consists of 14 features, namely Patient ID, Global number, Sex, Year of Birth, Age, Country, Province, City, Latitude, Longitude, Infection case, Infection order, State, and Label. In total, seven features are categorical, and the other seven features are string values. The dataset was pre-processed to convert all the string values into categorical values using the label encoder. The label encoder fits to transform the technique used to encode all the string values in the dataset in order to work with machine learning algorithms. The encoded features are sex, age, country, province, city, infection case, and state. The total number of datasets is around 160 cases, and target values contain two labels as 0 and 1. Class 0 corresponds to isolated, and class 1 corresponds to release. The correlation coefficient of the Pearson used in the dataset to explain the relationship between the characteristics of the element (independent values) and the target values (dependent values). The above figure provides a clear indication of the correlation between (-1, 0, 1). The below Equation (4) represents the Pearson’s correlation coefficient formula.

\[ r = \frac{(X - \bar{X})(Y - \bar{Y})}{\sqrt{(X - \bar{X})^2 \cdot (Y - \bar{Y})^2}} \]

If the correlation is -1, there exists a negative correlation where the features provide a negative relationship between the values of the function, and these features are discouraged. If the correlation is 1, it indicates the clear indication of the strong correlation between feature label values and correlation 0 shows absolutely no relationship between the feature labels. The dataset is divided into training and testing sets with 70% and 30%, respectively. The covid-19 dataset was analysed using the different algorithms mentioned in section 4. The algorithms compared with varying metrics of performance, such as accuracy, precision, recall, and f1-score, which are determined using the confusion matrix. From the analysis, it is clearly seen that Naïve Bayes algorithms outperform all the other algorithms, due to the number of data points is too small, and also it works independence features values. Table 1 represents the Overview of Machine learning Algorithms.
Table 1: Overview of Machine learning Algorithms

| Algorithms                  | Advantages                                                                 | Drawbacks                                                                 | Applications                                |
|-----------------------------|----------------------------------------------------------------------------|---------------------------------------------------------------------------|---------------------------------------------|
| **Naïve Bayes (NB)**        | • Less sensitive to missing data                                          | • Need to calculate the prior probability                                 | • News categorizing                         |
|                             | • Easy to understand the results                                         | • The error rate in classification results                                | • Email spam detection                      |
|                             | • Works well for small data                                               | • More susceptible to input form data                                     | • Face Recognition                          |
|                             |                                                                            |                                                                            | • Sentiment Analysis                        |
| **Decision tree (DT)**      | • Less time required for data pre-processing                             | • More complex compares to other algorithms                               | • Weather forecasting                       |
|                             | • No need for Normalization of data                                       | • Complexity in cost and time                                             | • Diabetic prediction                       |
| **Support Vector Machine (SVM)** | • Useful when we don't have an idea about data                          | • Choosing kernel function is not easy                                   | • Protein structure prediction              |
|                             | • Works with unstructured and semi-structured data                        | • Cost longer training time for more massive datasets                    | • Intrusion detection                       |
|                             |                                                                            |                                                                            | • Handwriting recognition                   |
| **Random forest (RF)**      | • Works well with high dimensional data without dimensionality reduction | • Leads to over fitting if noise presence in the data                     | • Outlier detection                         |
|                             | • Overcome over fitting problem                                          |                                                                            | • Clustering                                |
| **XGBoost (XGB)**           | • Better regularization                                                  | • Only work with numerical values                                        | • Weather forecasting                       |
|                             | • Parallel Processing                                                     | • Leads to over fitting when hyper parameters are not tuned properly    | • Diabetic prediction                       |
|                             | • Better handling missing Values                                          |                                                                            |                                             |
| **K-Nearest Neighbour (KNN)** | • Easy to implement                                                      | • High time complexity                                                   | • Classification                           |
|                             | • Easy to learn the non-linear boundary                                   | • Inefficient since entire data to be processed for every training       | • 3D structure of protein prediction        |
Other than Naïve Bayes, the XG boost algorithm outperforms all other algorithms due to the ensemble method. The overall accuracy varies from 74% to 100% for a different algorithm, which is pure, depends on the dataset samples. Figure 8 shows the Pearson Correlation mapping between the feature values. Figure 9 shows the Graphical View of performance analysis on a traditional ML algorithm. The precision measures exactness from the given data. Recall measures the percentage of relevant results that correctly classified by the algorithm. The F1-measure is the harmonic measure of recall and precision is shown in Table 2.

**Table 2: Performance analysis of traditional ML algorithm on COVID-19 data**

| Algorithm | ACC | PRE | REC | F1  |
|-----------|-----|-----|-----|-----|
| KNN       | 0.74| 0.74| 0.74| 0.74|
| NB        | 1   | 1   | 1   | 1   |
| DT        | 0.94| 0.96| 0.98| 0.97|
| RF        | 0.96| 0.97| 0.98| 0.98|
| SVM       | 0.92| 0.98| 0.92| 0.95|
| XGB       | 0.99| 0.98| 0.95| 0.99|

Fig. 8: Pearson Correlation mapping between the feature values

Fig. 9: Graphical View of performance analysis on a traditional ML algorithm

### 4. Conclusion and Future Scope

Corona disease state prediction from the Korean dataset on COVID-19 is analysed and classified using the traditional machine learning algorithms have implemented in this work. The attempt also provides a clear view of how the amount of dataset deciding the accuracy of machine learning algorithms. The XG boost algorithm outperforms all the algorithms in terms of accuracy, precision, recall, and F1-measure with 99%, 98%, 95%, and 99%, respectively. In the future, the algorithm can also make use to predict the release of the lockdown of the respective country, the spread of COVID-19, and the statistical model to network the affected zones and maximum zones that have been affected.

### References

[1]. Singhal, T. “A review of coronavirus disease-2019 (COVID-19),” The Indian Journal of Pediatrics, 1-6, 2020.

[2]. Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R, “COVID-19 infection: origin, transmission”, and characteristics of human coronaviruses. Journal of Advanced Research, 2020.

[3]. Liu, Y. C., Kuo, R. L., & Shih, S. R. “COVID-19: the First Documented Coronavirus Pandemic in History”,Biomedical Journal, 2020.

[4]. Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., & Lessler, J, “The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application.”,Annals of internal medicine, 2020.

[5]. Yan, L., Zhang, H., Goncalves, J. et al. “An interpretable mortality prediction model for COVID-19 patients.” Nat Mach Intell (https://doi.org/10.1038/s42256-020-0180-7, 2020.

[6]. Suriya, M., Chandran, V. and Sumithra, M.G., “Enhanced deep convolutional neural network for malarial parasite classification” International Journal of Computers and Applications, pp.1-10, 2019.

[7]. Suriya, M., Chandran, V. and Sumithra, M.G., “Design of Deep Convolutional Neural Network for Efficient Classification of Malaria Parasite”, 2nd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing (pp. 169-175). Springer, Cham, 2020.