Prediction and Diagnosis of COVID-19 Using Machine Learning Algorithms

Tejaswini K S [1], Nisarga Prasad B S[1], Nayana M[1], Jayashree B R[1], Shyamala S C[1]

[1]Student, Department of Electronics and Communication Engineering, P.E.S Institute of Technology and Management, Vishveshvaraya technological university, Belagavi, karnataka, India.

Corresponding Author: jayashribr9739@gmail.com

Abstract: - The world is reworking in a digital era. However, the field of medicine was quite repulsive to technology, recently, the advent of newer technologies like machine learning has catalyzed its adoption into healthcare. The blending of technology and medicine is facilitating a wealth of innovation that continues to improve lives. With the realm of possibility, machine learning is discovering various trends in a dataset and it is globally practiced in various medical conditions to predict the results, diagnose, analyze, treat, and recover. Machine Learning is aiding a lot to fight the battle against Covid-19. For instance, a face scanner that uses ML is used to detect whether a person has a fever or not. Similarly, the data from wearable technology like Apple Watch and Fitbit can be used to detect the changes in resting heart rate patterns which help in detecting coronavirus. According to a study by the Hindustan Times, the number of cases is rapidly increasing. Careful risk assessments should identify hotspots and clusters, and continued efforts should be made to further strengthen capacities to respond, especially at sub-national levels. The core public health measures for the Covid-19 response remain, rapidly detect, test, isolate, treat, and trace all contacts. The work presented in this paper represents the system that predicts the number of coronavirus cases in the upcoming days as well as the possibility of the infection in a particular person based on the symptoms. The ability of Science and Technology to improve human life is known to everyone and hence the use of technologies is increasing day by day. Machine Learning is one such field of technology that has become popular in a very short period of time.

Key Words: — ***Covid-19, Corona virus, India, Regression Model, Machine learning, Prediction, Diagnosis.***

I. INTRODUCTION

Corona viruses are a large family of viruses that are known to cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) [6]. These two diseases are spread by the corona viruses named as MERS-CoV and SARS-CoV. SARS was first seen in 2002 in China and MERS was first seen in 2012 in Saudi Arabia [8]. The latest virus seen in Wuhan, China is called SARS-COV-2 and it causes corona virus.

Corona virus spread from one city to whole country in just 30 days [50]. On Feb 11, it was named as COVID-19 by World Health Organization (WHO) [5].

Manuscript revised July 30, 2021; accepted July 31, 2021.
Date of publication August 02, 2021.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898

As this COVID-19 is spread from person to person, Artificial intelligence based electronic devices can play a pivotal role in preventing the spread of this virus. As the role of healthcare epidemiologists has expanded, the pervasiveness of electronic health data has expanded too [13]. The increasing availability of electronic health data presents a major opportunity in healthcare for both discoveries and practical applications to improve healthcare [48]. This data can be used for training machine learning algorithms to improve its decision-making in terms of predicting diseases.

As of May 16, 2020, totally 44,25,485 cases of COVID-19 have been registered and total number of deaths are 3,02,059 [3]. COVID-19 has spread across the globe with around 213 countries and territories affected [2]. As the rise in number of cases of infected corona virus quickly outnumbered the available medical resources in hospitals, resulted a substantial burden on the health care systems [44]. Due to the limited availability of resources at hospitals and the time delay for the

results of the medical tests, it is a typical situation for health workers to give proper medical treatment to the patients. As the number of cases to test for corona virus is increasing rapidly day by day, it is not possible to test due to the time and cost factors [25]. In our thesis, we would like to use machine learning techniques to predict the infection of corona virus in patients.

II. Personality Assessment Theory

During our research, we have investigated three algorithms through which we have performed supervised classification.

A. Support Vector Machines (SVM)

Support Vector Machines performs classification by constructing N-dimensional hyper plane that separates the data into two categories [12]. In SVM, the predictor variable is called an attribute and the transformed attribute is called a feature. Selecting the most suitable representative data is called feature selection. A set of feature describing one case is called a vector.

The ultimate goal of SVM modelling is to find the optimal hyper plane that separates the clusters where on one side of the plane there is target variable and on the other side of the plane other category. The vectors which are near the hyper plane are the support vectors [12]. In Figure, a typical example of support vector machine is depicted.

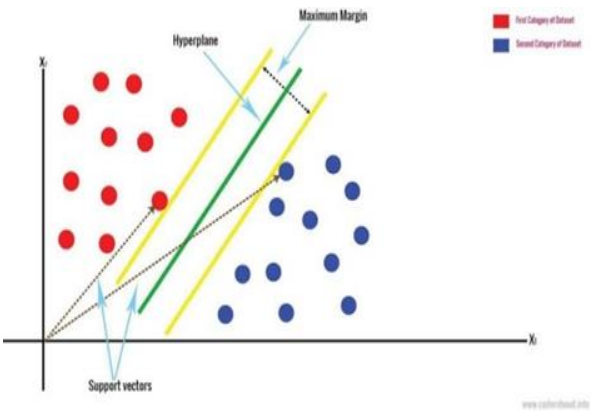

Fig.1. SVM Modeling

B. Artificial Neural Networks (ANN)

ANNs are an attempt, in the simplest way, to imitate the neural system of the human brain [53]. The basic unit of ANN are neurons. A neuron is said to perform functions on an input and produces an output [56]. Neurons combined together are called neural networks. Once the neural networks are formed, training data is started to minimize the error. In the end, an optimizing algorithm is used to further reduce the errors. The layered architecture of Artificial Neural Networks (ANNs) is represented in Figure.2.

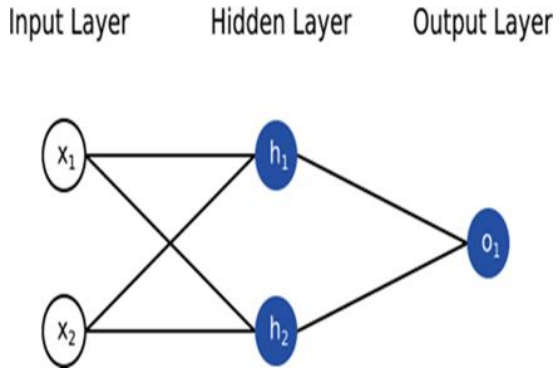

Fig.2. The layered architecture of ANN

C. Random Forests (RF)

The random sampling and ensemble strategies utilized in RF enable it to achieve ac- curate predictions as well as better generalizations [40]. The random forests consists of large number of trees. The higher the number of uncorrelated trees, the higher the accuracy [54]. Random Forest classifiers can help filling some missing values. Prediction in Random Forests (RFs) is represented in Figure.3.

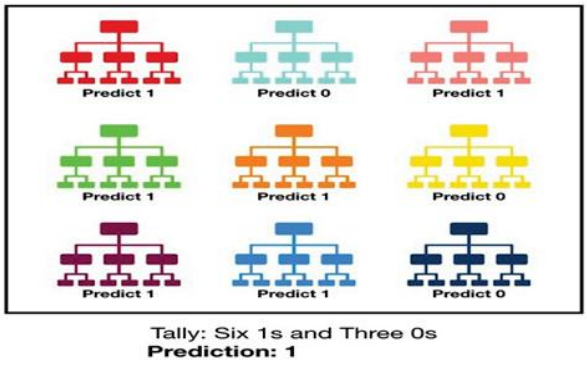

Fig.3. Prediction in Random Forests.

D. Implementation

The experiment was conducted in the Python IDLE, which is a default integrated development and learning environment for python. The experiment was conducted in various phases that are mentioned below:

- After data collection, the patient's data is divided into record sets containing 100 records, 150 records, 200

records, 250 records, 300 records, 355 records respectively.

- A 5-fold cross validation technique is used to randomize the testing data-set to get accurate results. Experiment on each machine learning algorithm is conducted by 5-fold cross validation with each of the record sets.
- The prediction accuracy of each algorithm at each record set is compared and evaluated for selecting the suitable algorithm for this data-set.
- A feature importance experiment is conducted to evaluate the importance of each attribute on the artificial classification task.

E. Algorithm Configurations

In this section, the configuration of the algorithms is mentioned. Changes made to the configuration of the algorithm can effect the results.

• *Support Vector Machines*:

SVC (kernel = 'linear', random_state = 0)

• *Artificial Neural Networks: Layers*:

ann.add(tf.keras.layers.Dense(units=6, activation='relu'))
ann.add(tf.keras.layers.Dense(units=6, activation='relu'))
ann.add(tf.keras.layers.Dense(units=1, activation='sigmoid'))

Compiling the ANN:

ann.compile(optimizer='adam',loss = 'binary_crossentropy', metrics = ['ac- curacy'])

• *Random Forests:*

RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state

= 0)

III. RESULTS AND DISCUSSION

This section presents the results that are obtained from the experiment. The performance metric mentioned in this Section is utilized to evaluate the performance of the algorithms that were selected after the Literature Review. Three algorithms that were identified as the most suitable for the classification task to predict COVID-19 are:

- SVM (Support Vector Machines).
- RF (Random Forests).
- ANN (Artificial Neural Networks).

Each of the above stated algorithms were trained with the data-set that was collected and results were interpreted. Performance of each algorithm was evaluated at different stages of training set. Each algorithm was trained with records sets containing 100 records, 150 records, 200 records, 250 records, 300 records, 355 records respectively. This experiment is performed to obtain which algorithm would be the most suitable for prediction of COVID-19. Also, as the data is split into smaller sets, we could also asses which algorithm would perform better with different datasets available.

A. Support Vector Machines (SVM) Results

Support Vector Machines (SVM) algorithm is trained with each record sets to identify its accuracy at all stages. At all stages, the data was divided into training and test data by using k-fold cross validation (5-folds). SVM achieves an accuracy of 98.33%.

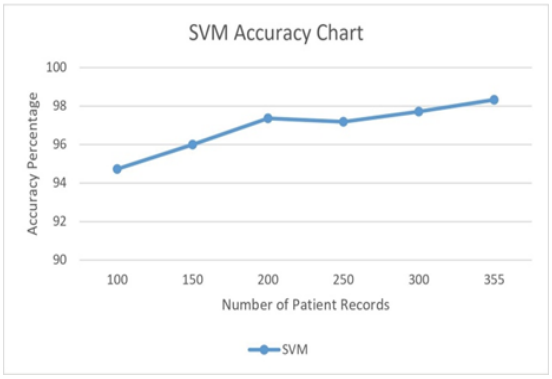

Fig.4. SVM Accuracy analysis

B. Random Forest (RF) Results

Random Forest (RF) algorithm is trained in a similar way with each records set to identify its accuracy at all stages. At all stages, the data was divided into training and test data by using k-fold cross validation (5-folds). RF achieves an accuracy of 99.44%.

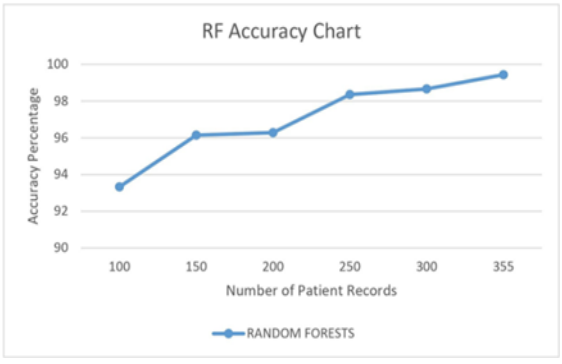

Fig.5. RF Accuracy Analysis

C. Artificial Neural Networks (ANN) Results

Artificial Neural Networks (ANN) Algorithm is trained on data with record sets and tested. On implementing ANN Algorithm, it achieves an classification accuracy of 99.25%.

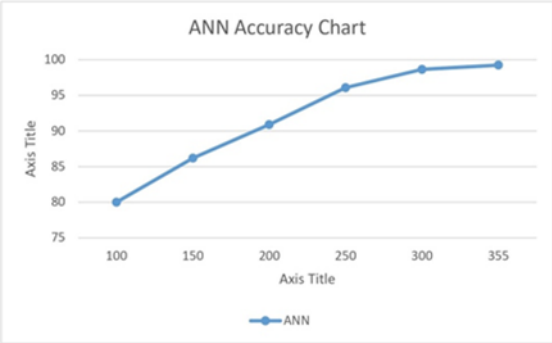

Fig.6. ANN Accuracy Analysis

D. Results Comparison

Based on the experiments conducted, the overall accuracy results are tabulated for comparison. A pictorial representation of performance of each algorithm at different record sets is presented in Figure.7.

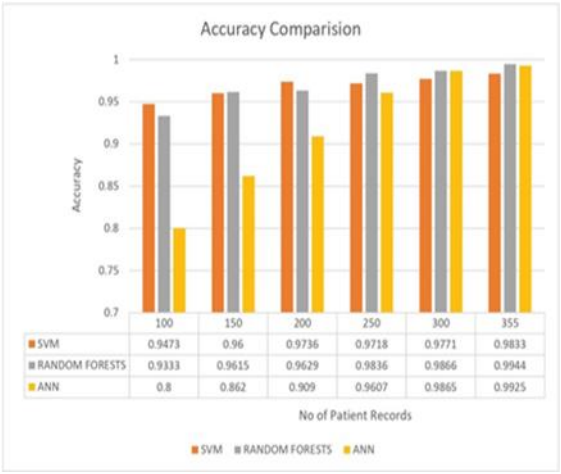

Fig.7. Performance of each algorithm at different record sets.

IV. Conclusion

In this research, a systematic literature review has been conducted to identify the suitable algorithm for prediction of COVID-19 in patients. There was no pure evidence found to summarize one algorithm as the suitable technique for prediction. Hence, a set of algorithms which include Support Vector Machines (SVM), Artificial Neural Networks (ANNs) and Random Forests (RF) were chosen. The selected algorithms were trained with the patient clinical information. To evaluate the accuracy of machine learning models, each algorithm is trained with record sets of varying number of patients. Using accuracy performance metric, the trained algorithms were assessed. After result analysis, Random Forest (RF) showed better prediction accuracy in comparison with both Support Vector Machines (SVM) and Artificial Neural Networks (ANNs). The trained algorithms were also assessed to find the features that affect the prediction of COVID-19 in patients.

There is a lot of scope for Machine Learning in Healthcare. For Future work, it is recommended to work on calibrated and ensemble methods that could resolve quirky problems faster with better outcomes than the existing algorithms. Also an AI-based application can be developed using various sensors and features to identify and help diagnose diseases.

As healthcare prediction is an essential field for future, A prediction system that could find the possibility of outbreak of novel diseases that could harm mankind through socio-economic and cultural factor consideration can be developed.

References

[1]. Coronavirus Disease (COVID-19) - events as they happen. Library Catalog: www.who.int.
[2]. Countries where Coronavirus has spread - Worldometer. Library Catalog: www.worldometers.info.
[3]. COVID-19 situation reports. Library Catalog: www.who.int.
[4]. Diagnosis of covid-19 and its clinical spectrum dataset.
[5]. WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020.
[6]. WHO EMRO | Questions and answers | COVID-19 | Health topics.
[7]. Support Vector Machine Machine learning algorithm with example and code, January 2019.
[8]. Ali Al-Hazmi. Challenges presented by MERS corona virus, and SARS corona virus to global health. Saudi journal of biological sciences, 23(4):507–511, 2016.
[9]. Sina F Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M Atkinson. Covid-19 outbreak prediction with machine learning.
[10]. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064–1069, 2016.