International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Machine Learning Based Approach For Corona Virus Disease Recovery Prediction

S. Kannimuthu[1], A. Arunkumar[2]

[1]Department of CSE, Karpagam College of Engineering, Coimbatore, India
[2]Department of CSE, Sri Krishna College of Engineering and Technology, Coimbatore, India

Abstract:- Coronavirus disease, aka COVID-19 is an transmittable disease caused by a newly discovered coronavirus. COVID-19 escalates largely over proximity with a diseased individual at the time of coughing or sneezing. When people touch their nose, eye or mouth, this disease spreads after touching a surface or anything that consists the virus on it. This virus can kill more than 30% of the infected person. Of late, this infectious disease causes the biggest burden on people all over the world. It also has a high death rate. Recently, machine learning based algorithms are being successfully exploited for classifying the data which are successfully adopted in many application areas. Machine learning methods are also can be applied to detect the COVID-19. Feature selection in Machine Learning process plays a significant role in improving the accuracy and other performance aspects. In this work, an approach for prognosing COVID-19 disease recovery is proposed which is realized to be the efficient method. Experimental result show that proposed model outperforms well when compared to other machine learning methods.
Keywords:- Corona Virus, Disease Prediction, Feature Selection, Machine Learning, Performance Evaluation.

1. INTRODUCTION

Coronavirus disease (COVID-19) is a transmittable / infectious disease caused by a new virus. The disease leads to severe breathing problem with symptoms such as fever, cough. People can protect themselves by washing their hands often, avoiding touching their face and evading close interaction with people who are suffered. COID-19 spreads mainly through with a person who suffered from this disease when they sneeze or cough. When people touch their nose, eye or mouth, this disease spreads after touching a surface or anything that consists the virus on it. This disease can be serious and even incurable. Older people and the person with other medical conditions seems to be more vulnerable to flattering severely ill. Presently no vaccine available to prevent corona virus disease [1].

1.1 Structure of a Corona Virus

The structure of a corona virus is presented in Figure1, Corona virus is in spherical shape which have protrusions and are crown-like. The corona virus has a diameter of 75-160nm and the virus genome is a continuous linear single-stranded Ribonucleic acid (RNA).

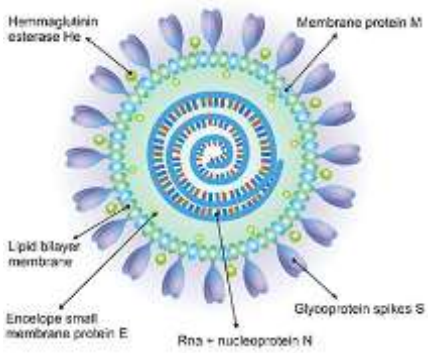


Figure 1. Structure of a Corona virus [2]

The genome of corona virus encodes a spike protein (S), an envelope protein, a membrane protein, and a nucleoprotein in this order. Spike protein is the utmost vital surface membrane protein of coronavirus among the above. Most corona virus are not dangerous but those that cause severe acute respiratory syndrome (SARS) can be deadly. The origin of corona virus is not clear. But experts say that originate in animals-like civets, bats and camels-and are usually not communicable to humans. Occasionally, coronavirus mutates and may spread from animals to humans and then from human to human [2].

Corona virus affected cases all over the world is exponentially increasing which leads to many deaths. So many people are not able to enjoy their day-to-day life because of this dangerous disease. Of late, artificial Intelligence (AI) techniques are mainly used in the diagnosis and treatment of patients who require care. Incredible amount of automation and technology plays a vital role in medicine. Some of the examples of AI already used in medicine are:

- Decision Making Systems and Assistance
- Surgeries done by a robot.
- Information Systems used in Laboratory
- Medical Therapy etc

In this work, a machine learning model is proposed to forecast the corona disease recovery of the people effectively and efficiently.

The significant contributions of this work are mentioned below:

1. A detailed Literature review on Corona Virus Disease is presented.
2. A Machine Learning Model for Corona Virus Disease Recovery Prediction is presented.
3. A detailed experimental analysis of numerous machine learning algorithms on various performance measures is done.

The remaining part of the paper is summarized as follows. Section 2 presents state-of-the art approaches pertaining to the COVID-19 research. Section 3 discusses about machine learning concepts. Section 4 describes about dataset information. Section 5 explores the method of research. Section 6 expresses comprehensive experimental evaluation and results. Conclusion is presented in Section 7.

2. RELATED WORKS

Jonathan et al [3] studied about the prediction of epidemic corona virus by fixing a deterministic SEIR meta-population transmission model considering human-to-human transmission alone. The estimate reports of SARS and MERS-CoV were lesser than that of the reproductive number of this coronavirus. The predicted estimation range was between 1.1

and 4.2 for SARS corona virus (predominantly from 2 to 3). MERS-CoV estimates had lower range that the ones that were collected from Saudi Arabia. The mean value was less than 1.0(approximately 0.5), however showed larger temporal inconsistency with increase in period of time especially in health care models. The airline travel was the only model that was taken into account for the spatial component. The local connectivity such as rail and road transportation were underrated. The assumption that the reproductive numbers had little heterogeneity may make a variation in the estimation of reproductive number. Furthermore, as case information increases, there is a reduction in RR0 estimation. A significant percentage of these infections had been because of multiple exposures to animals that were incomplete in some way or the other. These may symbolize a phase of high transmission that will not persist over long terms. The high transmission may be due to the stochastic distinction, complimentary seasonal conditions or selection bias in the observation of large clusters of transmission.

Zhou et al [4] proposed a methodology which is based on the proteins that regulate the viral infections which are restricted to the corresponding sub-network contained by widespread human interactome network. To make a drug work successfully for HCoV, the corresponding proteins must be contained in the on-the-spot surrounding area of the corresponding sub-network in the human protein-protein interactome. Network proximity measure was used for enumerating the association between subnetwork that is specific for HCoV and drug targets of the humans' molecular interactions in a particular cell. A drug-target network was established to gather target information for approximately 2000 experimental drugs approved by FDA. There was an incorporation of PPIs with five types of experimental data, such as, (1) binary PPIs from 3D protein structures; (2) binary PPIs from impartial high-throughput yeast-two-hybrid assays; (3) experimentally recognized kinase-substrate interactions; (4) signalling networks resulting from experimental data; and (5) literature-derived PPIs with various experimental indication for the betterment of quality and comprehensiveness of the human protein interactome network. Finally, to infer from the research of bias in the network proximity analyses, Z-score (Z) measure and permutation test were done.

Farid et al [5] grouped the statistical and machine learning tools to mine the features from CT images using four image filters along with composite hybrid feature extraction (CHFS). The chosen features were processed by the stack hybrid classification system (SHC) for classification process. MPEG7 Histogram Filter was used to depict the composition of an image. This histogram is very constructive for indexing and extracting pictures using different image classes. Gabor Image Filter was used for texture analysis which analyses whether or not there are any dissimilar frequency information in the picture in particular directions. The proposed model demonstrates better for optimal feature selection than the conventional classification approaches. This model efficiently cut down the false-negative rate with higher accuracy in association with Naïve Bayes as a meta-classifier in a hybrid classification method.

Fan et al [6] started a study for calculating the distribution and scale of travellers residing in Wuhan approximately. The study also analysed the social features of the travellers, such as, education level, reasons for migration, outline of migration and number of members from each family migrating. This may be valuable for medical professionals, virologists and epidemiologists to extend the research of the virus spread and advance the accuracy of prediction model. The data set used was based on China Migrants Dynamic Survey done during the year 2013–2018. The survey was accomplished through a stratified sampling method, and the mode of collection of data was structured questionnaires. The two parts of analysis were, an examination of the source of Wuhan's fluctuating population at the regional

level using past data, and a study of the moving population inside Hubei Area based on 2017 data.

Wu et al [7] studied the multiplicative number of 2019-nCoV based on the total number of cases reported and estimation calculation was done to find out the sum of cases exported outside mainland China. All through the period of a pandemic, the reported cases were increasing exponentially when human-to-human interaction was getting higher. The modelling techniques used for the study was similar to that of one used by other scientists who were working on same operations and targets. The model was parameterised with the latest mobility data from OAG. Identification and prevention of zoonotic source stays a significant task to eliminate new animal-to-human seeding events.

Malik et al [8] made an effort to disclose the evolutionary perspective of 2019-nCoV using the comprehensive genome analysis as the base. Maximum likelihood method (ML) in association with MEGA 7.0 was developed for phylogenetic analysis. The estimation of 2019-nCoV occurrence was done using MegAlign software of DNASTAR.

Tian-Mu et al [9] developed RP transmission model, which was used to fit the described data in Wuhan City. The main objective was to make available a mathematical model for the calculation of transmission of SARS-CoV-2. The R0 of SARS-CoV-2 was 3.58, which varied from person to person. Different types of transmission methods produced different values of R0. Many different methods can be employed to calculate the R0 based on the epidemic growth rate of the epidemic curve. The ratio of asymptomatic infection of MERS and SARS was below 10%.

Kuniya [10] employed the SEIR compartmental model on the reported cases of COVID-19 and found that the basic reproduction number R0 is 2.6. Also, predicted that the pandemic would top probably the middle summer. COIVD-19 will vanish in the summer is a false promise. This study did not alter the fundamental reproduction number R0 from the early estimations done by other researchers. The identification rate ‘p’ in a realistic parameter range 0.01 – 0.1 won’t have an effect on the vital epidemic size, characterized by R0. There was a positive consequence on the delay of the pandemic peak due to the interference. This could add some improvement in the medical setting. The impact of the epidemic will diminish if the interference lasts for moderately longer period. Lessening of the epidemic risk is directly proportional to the actual infective population.

Kian et al [11] demonstrated a model based on a study that the viruses that are exhibited to less cruel surroundings need a more severe encasement for defending their virions from getting harmed. One approach to determine the severity of encasement is by observing the level of predicted fundamental disorder in the capsid and matrix proteins. Higher disorder levels in the shell proteins are one of the factors for the rapid spread of many CoVs through respiratory mode. This may be due to the large PID levels of HCoVs in matrix and nucleocapsid proteins. The outcome of the disorder analysis can also be used to forecast the behaviour of CoV. Moreover, there is a close resemblance between SARS-CoV and PEDV in terms of their PID levels in the matrix and nucleocapsid proteins.

Nedialko et al [12] stated that the SIR models fail in modelling some significant views of the disease spread. While considering the outbreak of SARS during 2002 – 2003, the R0 estimation on taking into account the early outbreak of SARS was between 2.2 and 3.6. The limited spread of SARS was due to the strict isolations on infected persons forced by world public health agencies. SIR model considers the data to be a fully mixed, homogeneous population, where every person has same quantity of contact as others. Thus, SIR models could not precisely modeled the enlarged time of contact at places like hospitals and the diminished time of contact of isolated persons. The R0 estimation must have been more precise, if the population outside the hospital had equal number of contacts established to the

population within the hospital. Thus the SARS infection might have involved many people. The existing work has a high-fidelity model which makes use of agent-based simulations, in which every individual person is tracked as there is a movement. These models engage a difficult parameterization and involve extensive computation.

Ray et al [13] analysed the potential of ensemble methods to provide advancements in infectious disease predictions. These ensemble methods mainly concentrated on those methods which estimated the weighted averages of predictive distributions. This method attained a performance with an increased stability, which was approximately equal to the individual component models. CW and FW-reg-w models have better average performance to some extent at some point in the test phase than some ensemble methods that integrated some regularization on the weighting functions. As only one fourth of the data set has been used for modelling, it might have negative impact on the weighting functions. The ensemble method is appropriate for combining predictions from component model collections, where each yield a full predictive distribution, independent on model structure.

Zhan et al [14] developed a controlled non-linear programming, which outputs the best set of estimates for the unidentified states and parameters. This model also allows projecting the count of infected as well as exposed persons. Diverse outputs and conclusions may be obtained by the usage of different models of the transmission process. This study helps to have an estimation of the number of individuals affected by COVID-19 and predicts that COVId-19 may top in the mid of March-2020.

Machine Learning

Machine learning is the algorithmic method of data analysis that computing systems use to perform precise task without using obvious instructions, depend on inference and patterns. It is the branch of Artificial Intelligence (AI). Machine Learning algorithms constructs mathematical model or any other representations based on training samples with the intention of making predictions. It is used in many applications such as fraudulent detection, spam filtering, computer vision and disease predictions. It is infeasible to develop a basic algorithm for effectively performing the task.

Machine learning approaches are classified into

1) Supervised Learning : Learning happens with a labelled training set
2) Unsupervised Learning : Identify patterns without any class labels
3) Semi-supervised Learning : Combines small number of labelled samples and more amount of unlabelled samples
4) Reinforcement Learning : System learn to act based on reward/feedback [15,16]

Data Set Description

The dataset is collected from The *New York Times* [17] who is documenting confirmed Covid-19 cases at the county level. This may be the most gritty, detailed case dataset available to the public. The set of important attributes available in raw data is mentioned in Table 1.

Table 1. Details of confirmed Covid-19 patient

S. No	Attribute Name	Description
1.	Country	Country name of the patient
2.	Location	Location of the patient
3.	Summary	Patient details while admitted in the hospital
4.	Gender	Gender of the patient
5.	Age	Age of the patient
6.	Hosp_visit_date	Date of admission in the hospital

7.	Death	Information about whether patient is dead or not
8.	Recovered	Information about whether patient is recovered from disease or not
9.	Symptom	List of symptoms faced by the patient in a textual format

3. METHODS OF RESEARCH

The architecture of disease recovery predictor system is illustrated in Figure 2. In this work, raw data is collected from the source and represented that is convenient for training the model. The raw data has duplicates, errors and missing values. These issues are resolved in data pre-processing step. Data conversion and normalization task are also done in this step. The input vector is relevant for the data analytics are chosen by using feature selection step.
Before training, it is needed to choose an algorithm for evaluating the model. The main intension of training process is to make prediction correctly whenever needed. Every iteration of process is a training step. Using suitable performance measures, it is necessary to evaluate the performance of the various machine learning model and choose best algorithm for making prediction. It is done in evaluation step. Hyperparameter tuning is also done in this step. Using test data, the model is evaluated how the model will perform in the real world.

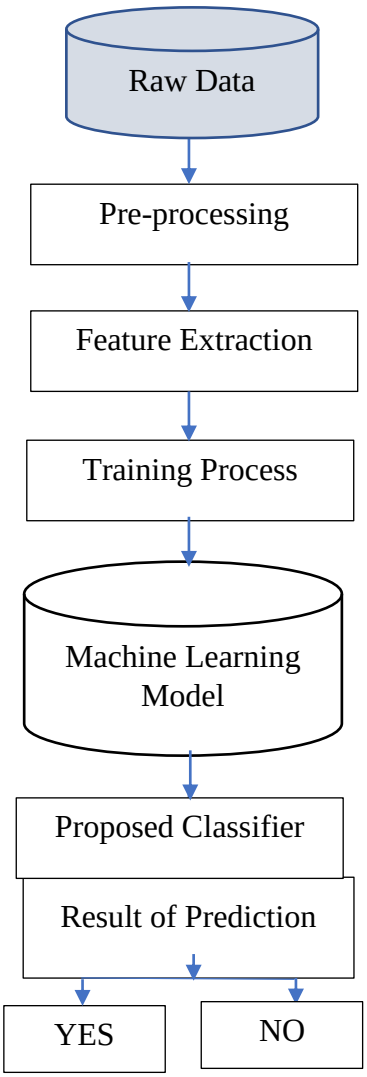


Figure. 2. Architecture of Disease Recovery Predictor

The description of the set of algorithms utilized for the experimentation is tabulated below

S. No	Algorithm	Description
1.	**Decision Tree (DT)**	One kind of supervised machine learning which is a simple representation for classifying the dataset. Here the data is continuously split based on a certain input vector. This algorithm works well for both categorical and numerical attributes.
2.	**Naïve Bayes (NB)**	The classification algorithm based on Bayes' theorem with the assumption of all input vectors are independent. Naïve Bayes approach assumes that existence of an input attribute in a class is no way related to the occurrence of any other attribute.
3.	**Logistic Regression (LR)**	This approach aims to find the relationship between input vectors and the probability of certain outcome. The underlying method is similar to Linear Regression.
4.	**Random Forest (RF)**	This algorithm is used for both classification and regression analysis. This algorithm creates the set of trees on datasets and gets the predictions from each of them and choose the best solution by means of voting method. It is the example for ensemble method (i.e., decision by committee)
5.	**Gradient Boosted Trees (GBT)**	It is the example for boosting algorithm. The primary objective of this function is to define a loss function and minimize it. The prediction model is generated such that the loss function is minimum. It uses gradient descent and update predictions based on a learning rate.
6.	**K-Nearest Neighbour (KNN)**	It is a simple algorithm that keeps all data samples and classifies the new samples based on a distance measures (i.e., Euclidean Distance). It is used in pattern recognition and statistical estimation. It is also called as lazy learner.

4. EXPERIMENTAL RESULTS

The raw data collected form The New York Time source is pre-processed applied to the machine learning algorithms for forecasting whether the patient affected by COVID-19 is able to recover or not. Machine Learning process is implemented using Python Language.
The pre-processed dataset consists set of relevant parameters which are listed in Table 2.

Table 2. Pre-processed dataset information

S. No	Name of the Attribute	Description
1.	Sex	Patient's gender {male, female}
2.	Age	Age of the patient {young, middle_aged, elderly person}
3.	Fever	Patient has Fever or not {Yes, No}
4.	Cough	Patient has Cough or not {Yes, No}

5.	Breathing Problem	Patient has Breathing Problem or not {Yes, No}
6.	Chills	Patient has Chillness or not {Yes, No}
7.	Joint Pain	Patient has Joint Pain or not {Yes, No}
8.	Throat Pain	Patient has Throat Pain or not {Yes, No}
9.	Pneumonia	Patient has Pneumonia or not {Yes, No}
10.	Headache	Patient has Headache or not {Yes, No}
11.	Death	Information about whether patient is dead or not
12.	Recovered	Information about whether patient is recovered from disease or not

In this experimentation, Gender, Age and symptoms such as Fever, Cough, breathing problem, Chills, Joint pain, Throat pain, Pneumonia and Headache are considered for generating machine learning model. The attribute 'Recovered' is considered to be a class label (i.e., Target Attribute). In this experimentation, performance parameters such as accuracy, classifier error, kappa, weighted mean recall, weighted mean precision, absolute error and root mean squared error are considered. Table 3 show the record of performance evaluation of machine learning approaches on various measures.

Table 3. Record of Performance evaluation of Machine Learning
Algorithms on various measures

Performance Measures	**Decision Tree**	**Naïve Bayes**	**Logistic Regression**	**Random Forest**	**Gradient Boosted Trees**	**K-Nearest Neighbour**
Accuracy	83.65	86.76	86.91	85.83	86.55	87.63
Classification Error	16.35	13.24	13.09	14.17	13.45	12.37
Kappa	0.059	0.02	0.016	0.049	0.101	0.199
Weighted Mean Recall	61.22	63.1	59	58.94	62.77	70.24
Weighted Mean Precision	65.55	62.2	63.64	69	71.58	84.7
Absolute Error	0.256	0.259	0.26	0.242	0.263	0.243
Root Mean Squared Error	0.387	0.368	0.369	0.362	0.365	0.353

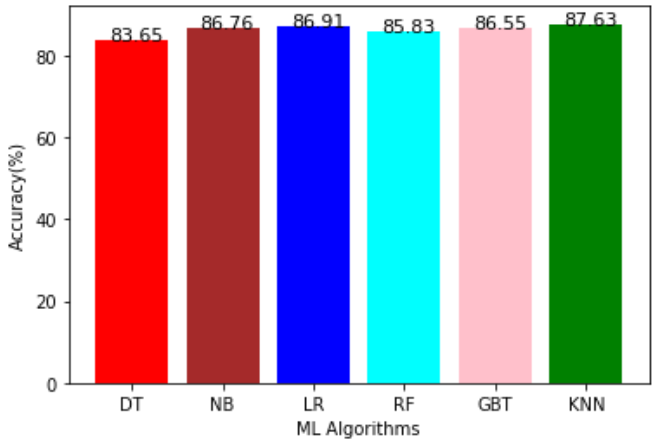


Figure 3. Comparison of ML Algorithms in in terms of Accuracy

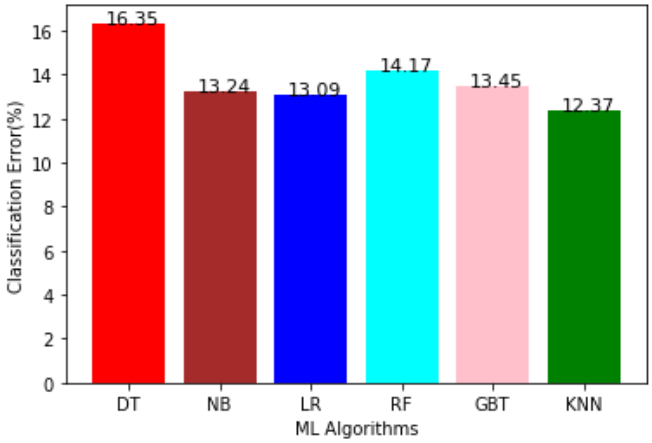


Figure 4. Comparison of ML Algorithms in in terms of Classification Error

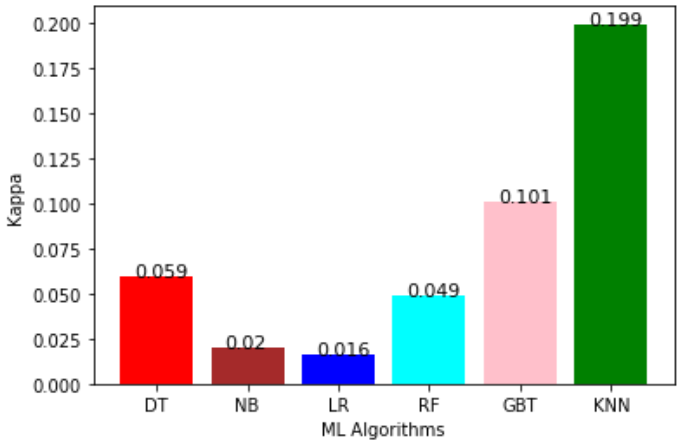
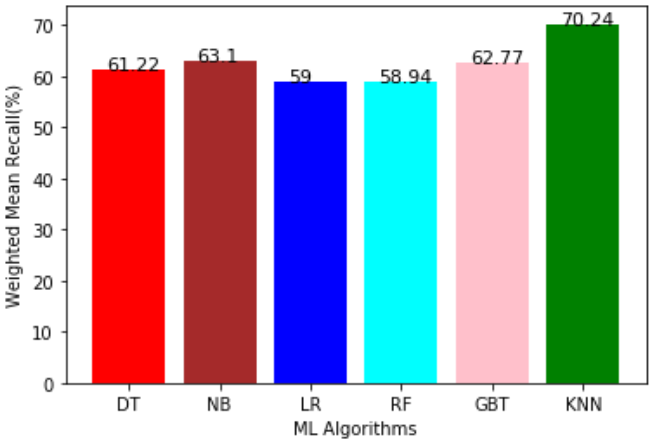


Figure 5. Comparison of ML Algorithms in in terms of Kappa



6

Figure 6. Comparison of ML Algorithms in in terms of Weighted Mean Recall

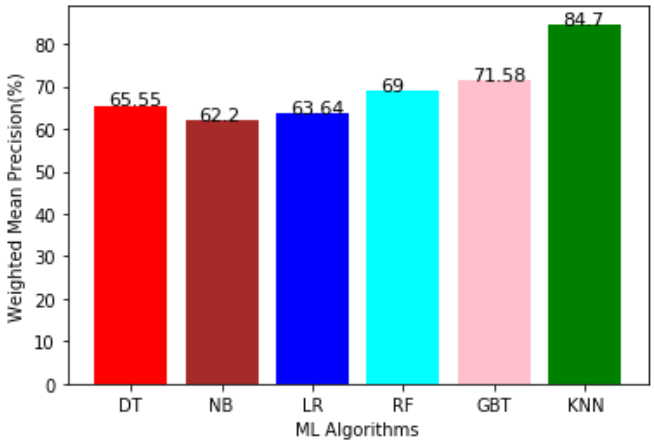


Figure 7. Comparison of ML Algorithms in in terms of Weighted Mean Precision

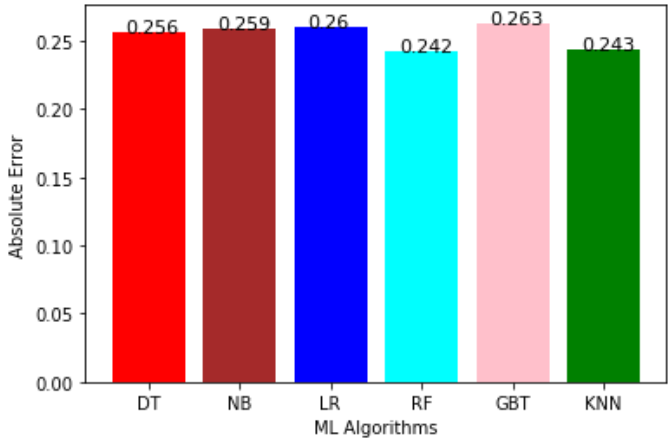


Figure 8. Comparison of ML Algorithms in in terms of Absolute Error

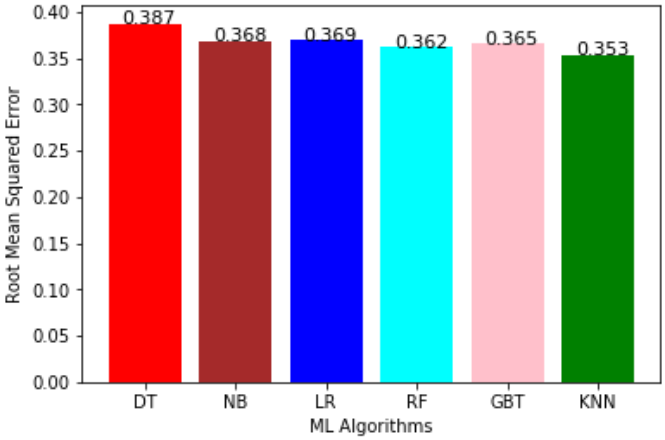


Figure 9. Comparison of ML Algorithms in in terms of RMSE

From Figure 3, it is clearly understood that K-Nearest Neighbour algorithm outperforms well in terms of accuracy measure when compared to other machine learning algorithms. Hence classifier error is less than other approaches which is illustrated in Figure 4. The K-NN algorithm predicts data with unknown label with high precision and recall when compared to other approaches. This is illustrated in Figure 6 and Figure 7 respectively. Absolute error and Root Mean Squared Error values are also convincing in K-NN when compared to other machine learning algorithms (Refer Figure 8 and 9).

4. CONCLUSION

In this work, A machine learning based intelligent model is proposed to predict whether is able to recover from the COVID-19 disease. A comprehensive experimental evaluation and result are obtained by considering various performance measures. It is observed that K-Nearest Neighbour Algorithm outperforms well when compared to the other ML approach in terms of Accuracy, Classification Error, Weighted Mean Precision, Weighted Mean Recall, Absolute Error and RMSE.

5. REFERENCES

[1] https://www.google.com/covid19/
[2] https://www.cusabio.com/2019-novel-coronavirus.html
[3] Jonathan M. Read, Jessica R.E. Bridgen, Derek A.T. Cummings, Antonia Ho, Chris P. Jewell "Novel corona virus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions", medRxiv, the preprint server for health sciences, https://doi.org/10.1101/2020.01.23.20018549, January 2020.
[4]. Yadi Zhou, Yuan Hou, Jiayu Shen, Yin Huang, William Martin and Feixiong Cheng, "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2", Cell Discovery, Vol. 6, No. 14, 2020.
[5] Ahmed Abdullah Farid, Gamal Ibrahim Selim and Hatem Awad A. Khater, "A Novel Approach of CT Images Feature Analysis and Prediction to Screen for Corona Virus Disease (COVID-19)", Preprints (www.preprints.org), doi:10.20944/preprints202003.0284.v1, March 2020.
[6] Changyu Fan, Linping Liu, Wei Guo, Anuo Yang, Chenchen Ye, Maitixirepu Jilili, Meina Ren, Peng Xu, Hexing Long and Yufan Wang, "Prediction of Epidemic Spread of the 2019 Novel Corona virus Driven by Spring Festival Transportation in China: A Population-Based Study", International Journal of Environmental Research and Public Health, Vol. 17, No. 5, 2020.
[7] Joseph T Wu, Kathy Leung, Gabriel M Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study", The Lancet, Vol. 395, No. 10225, pp. 689-697, 2020.
[8] Yashpal Singh Malik, Shubhankar Sircar, Sudipta Bhat, Khan Sharun, Kuldeep Dhama, Maryam Dadar, Ruchi Tiwari, Wanpen Chaicumpa, "Emerging novel coronavirus (2019-nCoV) - current scenario, evolutionary perspective based on genome analysis and recent developments", Pubmed Journal, Vol. 40, No. 1, pp. 68-76, 2020.
[9] Tian-Mu Chen, Jia Rui, Qiu-Peng Wang, Ze-Yu Zhao, Jing-An Cui and Ling Yin, "A mathematical model for simulating the phase-based transmissibility of a novel coronavirus", Infectious Diseases of Poverty, Vol. 9, No. 24, 2020.
[10] Toshikazu Kuniya, "Prediction of the Epidemic Peak of Coronavirus Disease in Japan", Journal of Clinical Medicine, *J. Clin. Med, Vol. 9, No.* 3, pp. 789, 2020.
[11] Gerard Kian-Meng Goh, A. Keith Dunker, Vladimir N. Uversky, "Understanding Viral Transmission Behavior via Protein Intrinsic Disorder Prediction: Coronaviruses", Journal of Pathogens, Vol. 2012, No. 738590, 2012.
[12] Nedialko B. Dimitrov, Lauren Ancel Meyers, "Mathematical Approaches to Infectious Disease Prediction and Control", Tutorials in Operational Research Informs, doi: 10.1287/educ.1100.0075, 2010.
[13] Evan L. Ray , Nicholas G. Reich, "Prediction of infectious disease epidemics via weighted density ensembles", PLOS Computational Biology, Vol. 14, No. 2, 2018.

[14] Choujun Zhan, Chi K. Tse, Yuxia Fu, Zhikang Lai, Haijun Zhang, "Modeling and Prediction of the 2019 Coronavirus Disease Spreading in China Incorporating Human Migration Data", medRxiv, the preprint server for health sciences, https://doi.org/10.1101/2020.02.18.20024570, February 2020.
[15]. Stephan Marsland, "Machine Learning: An Algorithmic Perspective", Second Edition (Chapman & Hall/Crc Machine Learning & Pattern Recognition), 2014.
[16]. Tom M. Mitchell, "Machine Learning: A Guide to Current Research", McGraw Hill publishers, 1997.
[17]. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html