Received 15 December 2020; Accepted 05 January 2021.

An Improved Random Forest Algorithm for Predicting the COVID-19 Pandemic Patient Health

Sanjeev Kumar Sharma[1], Umesh Kumar Lilhore[2], *Sarita Simaiya[3] , Naresh Kumar Trivedi[4]
[1, 2,3,4] Chitkara University Institute of Engineering and Technology,
Chitkara University, Punjab, India
[1]sanjeevk.sharma@chitkara.edu.in, [2]umesh.lilhore@chitkara.edu.in, [3]sarita.simaiya@chitkara.edu.in, [4]nareshk.trivedi@chitkara.edu.in
Corresponding Author: * Sarita Simaiya
Email id: *sarita.simaiya@chitkara.edu.in*

Abstract: The 2019 pandemic of the Corona infectious diseases also called "COVID-19" has been emanated in the Asian country China in the year 2019. This is precariously unjustly preventing almost everything in human society. This same rapid pace as well as an incremental increase in the proportion of patient populations, consequently, required an efficient and accurate forecast of such an infected client's potential result for proper care utilizing the machine learning model. To overcome these needs, the research experiment strives to optimize Covid-19 patient's potential to estimate a revolutionary model focused on an improved random forest (IRF) methodology with safety various features. This proposed IRF method has been implemented to predict Covid-19 information about a person's condition with high-dimensional, unstable functionality. Initially, a random forest algorithm will be used for organizing the value of the dependent variable and reducing the measurements. By using the COVID-19 Kaggle online dataset set of data, the proposed IRF method as well as existing Random Forest (RF) and Support Vector Machines (SVM) has been implemented over the WEKA Machine learning simulation tool.

Keywords- *Random Forest, Covid-19, Pandemic, Improved Random Forest, SVM, Pandemic Patient Health*

1. Introduction

The Coronavirus issue has become an extremely contagious viral infection which has seized massive global international attention. Simulation of certain disorders may be highly relevant for their influence forecasting. Although classical, mathematical, forecasting might provide some adequate models, it can sometimes struggle to understand some complexities throughout the results. The Novel corona-virus has been mainly related to the family of SARS-CoV-2 "Severe acute respiratory syndrome corona-virus-2", also known as COVID-2019(corona-virus disease-19) emerged in the year Dec 2019 in China (capital Wuhan) through that became spread very quickly throughout the world. In some of the countries, these infectious diseases are presenting no or little markings, inpatient, during the very first five weeks of virus infection [1].

The frequency of devastation as well as disease sparked only by virus meant that now the WHO declared it just a worldwide pandemic. There seem to be now over 180 nations in the world affected by COVID-19 well according to the Centers for Disease Control including preventative measures; almost 51 states in the United States had also reported cases [2]. Now COVID-19 doesn't have any approved vaccinations as well as extroverts and introverts-viral. Earlier 2002 and 2003 SARS pandemic had also been managed as well as ultimately ended through standard prevention mechanisms, which include immigration bans & separation of patients. Similar strategies have been being implemented in approximately most of the COVID-19-infected counties.

Family and community propagation all over the World has been the viruses' common format of transmission. Consequently, this has become vital to evaluate its efficacy of an important way of restricting social interaction utilized throughout multiple levels of government. Sufficient statistics have currently accessible through community rate and include a reasonable measure regarding the feasibility of governmental guidance released at certain levels. This helps in creating a sort of 'private distancing' through restricting the maximum numbers of assemblies of 20 or lesser people [3].

Coronavirus does seem to be a virus reported as that of the source of just a pandemic of respiratory illness initially found in Wuhan, China. Soon throughout the endemic problem, plenty of the victims in Wuhan, China seemed to have some link to wide market selling seafood as well as pets, implying a transfer through commodity to individual. Even then, a majority of patients that have not had exposure to pet markets have been responsible; signifying that there would be trying to spread from human to human. It's uncertain during this moment how effectively or environmentally sustainable the whole virus spreads among people-CDCs [4, 5].

To fulfill this need, a novel strategy focused on an improved random forest algorithm aims to strengthen the estimation of Covid-19 infection control. The proposed improved random forest algorithm has been used to predict future relevant information with high-dimensional unstable features of all of Covid-19 patient health. The whole research study has been separated into two parts that cover primarily the introduction section, the related analysis, and secondly the implementations of proposed Improved random forest model and eventually the findings of simulation and assessment of performance.

2. Related Work

Covid-19 seems to be a popular area of research and has been carried by different researchers. Very few of those research has selected chosen are just as below-

Throughout the experimental paper [6] researchers selected how a random forest algorithm can be trained to estimate future rapid growth of residues when a SIP request been released throughout a region. This random forest has indeed been used since this disease forecast has indeed been demonstrated to become the most reliable. The key measures factors throughout estimating their steady growth have been elevation, wavelength, including advancement for each metropolitan area once a SIP policy has been implemented across a region.

Throughout the experimental paper [7] an MLP implied by the investigator was focused on computing neurons quantities in some kind of the following frame as its triggered description of measured neural network in either a preceding layer, associated only with the neuron. Activity applies to the quantities of input vectors that are used as references from the so-called hidden layers, which sometimes explicitly mapping the output variable (identification trigger), under other limitations (fractal dimension), however, masks it by eliminating some of the unnecessary variables. Throughout the analysis paper [8] authors identified those nations around the world responding to just the COVID-19 disease outbreak with even a variety of policies designed to minimize effects on the economy, they conclude for low- and middle-income communities would favor the production of physical investment across income and losses.

Throughout the research work [9] Corona-virus (also known as COVID-19) seems to be an extremely infectious disease that may have caught the widespread people's attention. Identifying these diseases can indeed be tremendously beneficial throughout their influence forecasting. Although traditional, mathematical, simulation provides generated significant, this can also refuse to understand the complexities throughout the results. Researchers are using a widely accessible database throughout this article, giving data about contaminated, recovered, and death certificates throughout 5 weeks from January to March 2020. The whole data set [10], which is designed as a regression model, will indeed be converted to a correlation data set being used for the formation of the artificial neural network (ANN) multi-layer Perception [11].

Inside an investigation paper[12] implemented linear regression, multi-layer accessibility to information as well as neuron auto-regression technique besides COVID-19 Kaggle statistics were used to predict this same observational instance with COVID-2019 disease as well as speed in India. The Anticipated shows how possible COVID-19 influence trends throughout India depending on data collected from Kaggle [13]. Now with widely accepted confirmation information, death, as well as retrieved instances along with all India, enable throughout predicting as well as predicting a not-so-distant future surrounded by white time. The task for the government and selection of data must also be insistently retained with potential evaluation or possible insight.

The Corona Virus Disease-2019 is originally referred to as just the Corona-virus gene around the research paper [14]. A vaccine-free virus has been going to create an uncertain hazard to people's life as well as business and monetary structures in every nation on the planet. Precariously, everything else in life becomes ruthlessly stopped. Various classification methods have been created utilizing machine learning models as well as their efficiency was computed as well as analyzed. The study paper [15] suggested a new method based on the "Fine-Tuned Random Forest" framework. This proposed method mainly enhanced using the features of the SVM classifier.

The main measurement susceptibility with West Nile-CoV-2 signaling seems to be a primary output criterion of testing virus identification experiments in the literature review [16]. Besides nine West Nile virus-CoV-2 assessments, observational transmitted signals were evaluated utilizing different concentrations throughout consolidated clinician substance measured by virtual Transcription droplet. Those other findings may notify the selection of tests for something like the management of the existing COVID-19 pandemic. The main focus in the studies paper [17] seems to be on the random forests suggested as in the 2000s as well by Leo Breiman to construct another predictive entire ensemble with anything other than a set of decision-making bodies that grow instead in randomly selected metadata feature space. Most of the attribute values of random forests are already observed, as well as the computational strength which drives a method seems to have little awareness, notwithstanding rising popularity and practical usage. Throughout this article, scientists provided a detailed assessment of both the Breiman (2004) random forest design that's quite similar only to the method.

Overall assumptions are obtained throughout the study paper [18] through gathering over the entire ensemble. As these specific components of its ensemble constitute tree-structured predictive, and then when every one of these structures was built just use an infusion in unpredictability, such techniques were named "random forests. Throughout the experimental paper [19] several capabilities of RF were addressed, despite evaluating possible shortcomings as well as promoting additional experiments regarding similarities of the whole methodology with the other popular approaches to the classifier, starting in the initial estimation of advancement through MCI towards AD. Random Forest (RF) method has been widely implemented presently in several science domains to every large feature but intra-source results. The objective was to investigate the latest technology by using RF through diagnosis by Alzheimer's disease based on multi-modal endocrinology-imaging data [20].

3. Materials & Methods

This section of the research covers the functionalities of proposed method and as well material, data descriptions use in the research.

3.1 Dataset Description

The database was collected from Kaggle online application data set named as "Novel Corona Virus 2019 datasets"[21]. The data set mainly contains the pre and post health care details of Covid-19 patient.

3.2 Random Forest Method

The Random Forest (RF) classification method was firstly suggested by Breiman in the year 2001. It is a supervised learning based classification method. Preferably, the Random Forest module produces two significant types of information: the measurement of its significance including its response variable, and sometimes a measurement including its internal data model (the similarity of various internal and external sources with each other) [22].

Phases in Random Forest Classification Algorithm:

1) **Determine a trained model:** This is the 1st phase of RF method which mainly determines the training model for classifier. A sample technique is utilizes in the collection of classification models from the original dataset.
2) **Built Random Forest structure:** This phase is the second phase of RF method which mainly responsible for construction of Random forest structure. This phase collect the data from its all bootstrap and generates n trees. Theses generated tress help in creation of RF.
3) **Voting phase:** This is the final phase of RF method which is basically deal with the voting or pooling phase. This phase help in determining correct and incorrect features for the each of the tree in forest.

3.3 Proposed Improved Random Forest Method

The Proposed Improved Random Forest method is an enhanced form of existing Random Forest machine learning classification method. Initially, a random forest algorithm will be used for organizing the value of the dependent variable and reducing the measurements. Next, through using a random forest model, both the specifically chosen basic characteristics have been utilized, as well as the F-measure factors have been determined as probabilities within each feature selection to construct the forecasting models for consumer health.

The IRF method mainly utilizes the feature of well known methods “bootstrap resembling”. Proposed method IRF method separates sub-groups from its samples ranged and generates tree structure for each sample. Furthermore, this same algorithm categorizes its tree structure but also utilizes a direct vote, again with the categorization and prediction.

3.4 Key Element in Proposed Improved Random Forest

The proposed IRF methods have following key features which help in analysis of Covid-19 statistics.

3.4.1 Feature Ranking:

The Proposed IRF method utilizes the ranking phases which assign the ranking for all the collected features. In this phase we use a “module slicing entropy feature”, which determine the preliminary feature and augmented a random list.

3.4.2 Determine the number of trees

The IRF method determines the total number of trees that will be added in the forest. Here we determine the key variables that regulate the efficiency of random forest classification. These key variables include association and precision. The consistency of identification becomes established according to intensity as well as comparison. Here we received various trees that are generated from the last phase, and now theses trees will be utilized in classifier performance formulations.

3.5 Proposed Improved Random Forest Model

Proposed IRF method is based on following steps:

i. *By utilizing various existing wide range functions construct patterns.*

ii. *Calculate all the values which include the significance rating. Firstly the existing number of characteristics and the second includes sorting of the entire component in decreasing order.*

iii. *Evaluate each extraction ratio as well as eliminate the insignificant index from its existing factor to get the latest collection of functions.*

iv. *Evaluate the tree further and calculate the percentage for each incorporate portion.*

v. *Implement feature ranking on leaves of trees. Compute the entropy for the tree by using the equation-*

$$En = \sum V_0 \log (n) / V_0(C) \qquad \text{----------------------------}(1)$$

Where: N= Node, T= Tree, $V_0(c)$ = Probability value for C,C= Class label and En= Entropy

vi. *Repeat steps from 1 to 5, till the final forest not constructed.*

4. Simulation Results and Comparisons

The proposed IRF method as well as existing Random Forest (RF) and Support Vector Machines (SVM) have been implemented over the WEKA Machine learning simulation tool. The proposed method firstly identified all the available healthcare data which includes the details related to number of hospitals, ventilators, beds, doctors and other essential equipments i.e. oxygen cylinder, medicines. This proposed method categories these health care data sets in to three categories which include: total number of health care objects, currently used and available.

The proposed IRF method also determines the features form Covid-19 dataset which includes, total infected patients and patient conditions. The IRF method divides the patient condition into two categories normal and critical. The figure 1.1 is showing role of proposed IRF method in the mapping of health care facility and Covid-19 patients. The proposed method helps in mapping of patient and heath care facilities.

Received 15 December 2020; Accepted 05 January 2021.

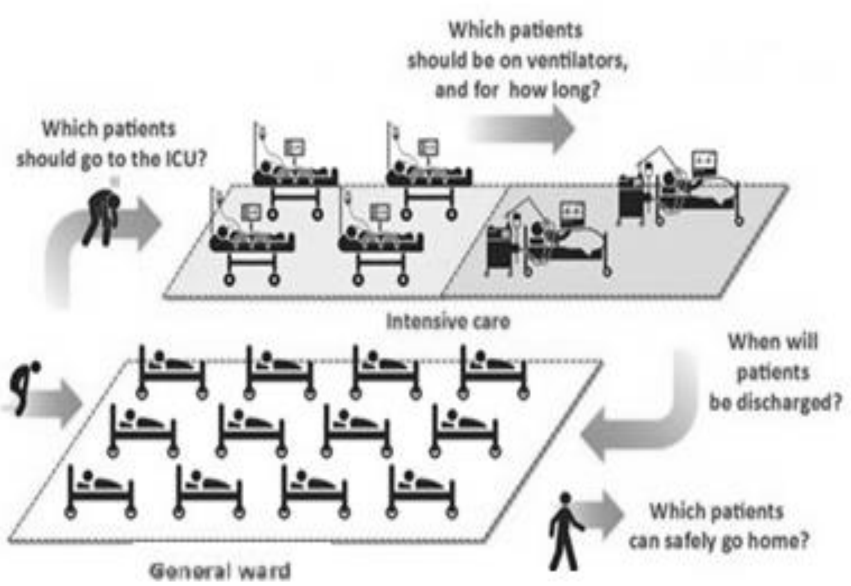
Which patients should be on ventilators, and for how long?
Which patients should go to the ICU?
Intensive care
When will patients be discharged?
Which patients can safely go home?
General ward

Figure 1.1 Mapping of Health care facility and Covid-19 patients [2]

4.1 Classifiers Performance Evaluation Parameters

To determine the performance of proposed IRF method following parameters [23-25] were used-

1) **Confusion Matrix –** This matrix defines the summary of prediction outcomes for a particular machine learning problem.
2) **True Positive Rate (TP) %-** It is also known as 'Sensitivity'. It is the ratio of true positive and true and false positive values.
3) **False Positive Rate (FP %)-** This is a ratio of the number of incorrect positive forecast and the number of negative objects.
4) **Precision %-** This is the fraction of all the relevant values which are correctly retrieved.
5) **Recall %-** This is the ratio of true positive contests from all possible positive contests.
6) **F-Measure (F1-Score) %-** This is a measure of experiments accuracy. It can be defined as the weighted harmonic mean of the experiments precision values and recall values.
7) **MCC %-** It is known as Matthews's correlation coefficient. It is mainly used to evaluate the performance between two two-classes (binary).
8) **ROC Area %-** It is known as Receiver Operating Characteristic. It is basically used to evaluate the efficiency level of a classifier (binary) by constructing a graph between true positive and false positive outcomes.
9) **PRC Area %-** It is also known as Precision-Recall Curve. It is a graph between accuracy plot value and recall of all major test results.

4.2 Result Comparisons:

This simulation was performed over a total of 30,200 COVID-19 patient data, by using WEKA machine learning tool. The proposed Improved Random Forest method and existing machine learning classifier method Random Forest and SVM has been implemented and tested for Covid-19 dataset.

The table 1.1 is showing a confusion matrix for the proposed IRF method for 30,200 COVID-19 patient data. In this table variable Co represents false and C1 represents True values. In the confusion matrix we can see the proposed IRF achieves 19,200 records for TN, 4,800 for FP, 4,200 for FN, and finally 12,000 for TP. The details over view of attributes are representing by figure 1.2, for Covid-19 patient datasets.

N=30,200	**a**	**b**
a=C0	TN=19200	FP= 4800
b=C1	FN= 4200	TP 12000

Table 1.1 Confusion Matrix for the IRF

Where C0= FALSE, C1= TRUE

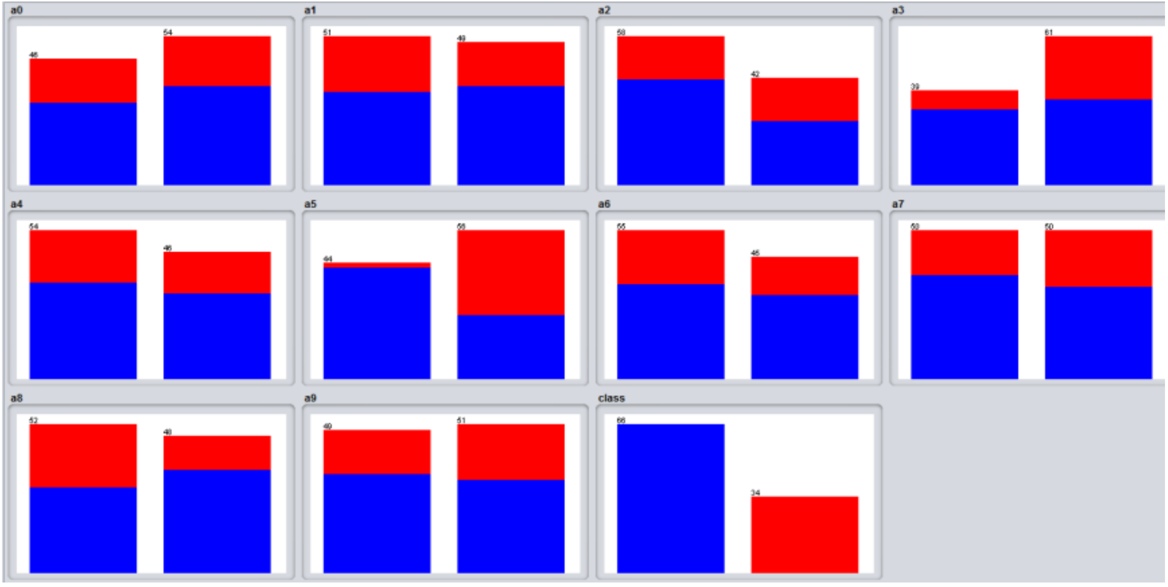

Figure 1.2 Attributes view of Covid-19 dataset

For Covid-19 patient health care data set a random forest has been constructed (as shown in figure 1.3). In this OUTPUR tree the variable C0 is representing the values that are classified as false and similar C1 representing the values which are classified as true.

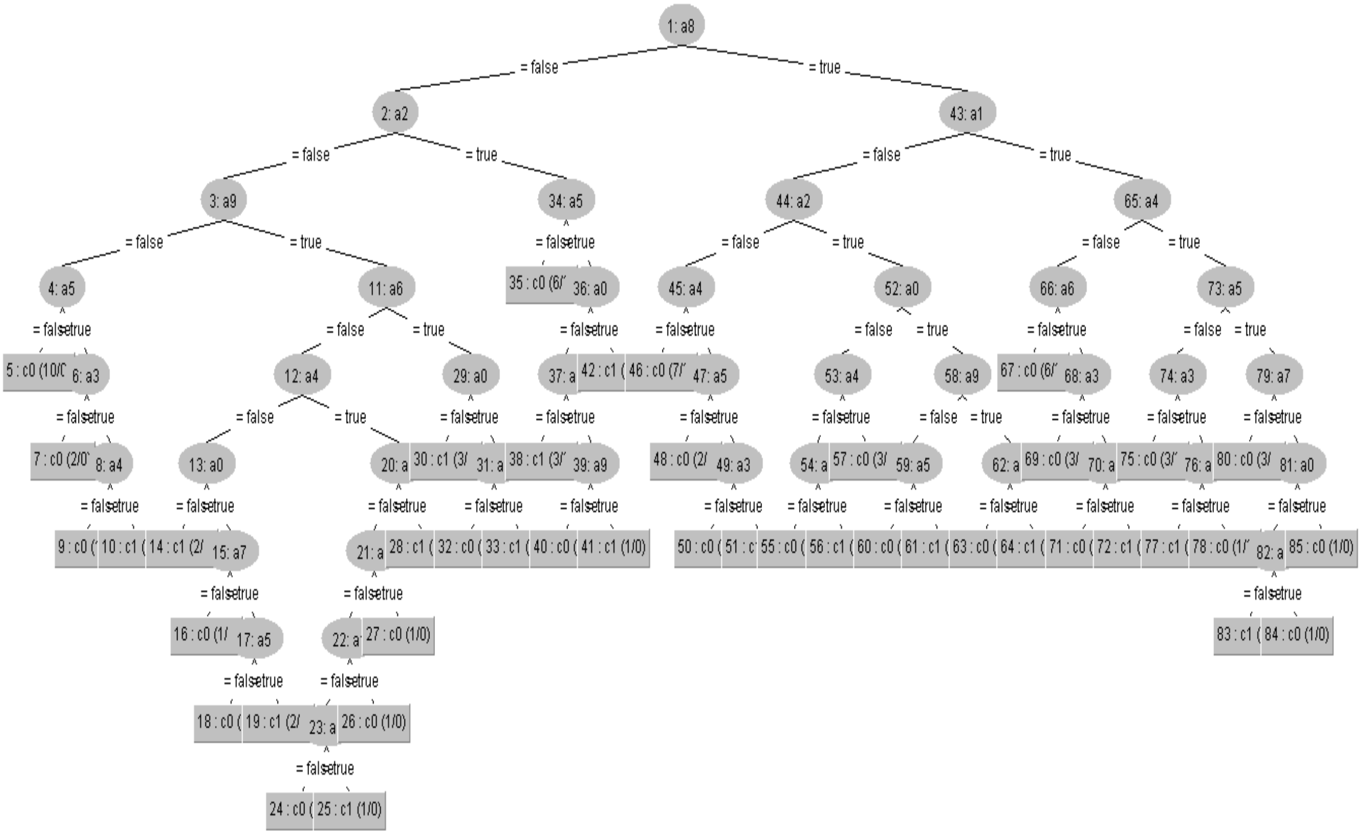

Figure 1.3 Tree view of proposed improved random forest method

Method	Class	TP Rate %	FP Rate %	Precision %	Recall %	F-Meas-ure %	MCC %	ROC Area %	PRC Area %
Proposed IRF	C0	0.8	0.286	0.8	0.8	0.8	0.514	0.757	0.758
	C1	0.714	0.2	0.714	0.714	0.714	0.514	0.757	0.628
	Weighted Avg.	0.765	0.25	0.765	0.765	0.765	0.514	0.757	0.704
Existing RF	C0	0.746	0.251	0.746	0.8	0.746	0.447	0.648	0.651
	C1	0.648	0.154	0.648	0.714	0.648	0.447	0.648	0.584
	Weighted Avg.	0.712	0.221	0.712	0.765	0.712	0.447	0.648	0.601
Existing SVM	C0	0.701	0.224	0.701	701	0.701	0.406	0.599	0.601
	C1	0.689	0.129	0.689	0.689	0.689	0.406	0.599	0.521
	Weighted Avg.	0.601	0.202	0.601	0.601	0.601	0.406	0.599	0.578

Table 1.2 Simulation result for Proposed IRF Vs RF Vs SVM Method

The WEKA simulation results for proposed IRF and existing RP, SVM has been displayed in table 1.2. Various performances measuring parameters i.e. TP %, FP %, Precision %, Recall %, F-Measure %, MCC %, ROC % and PRC %.

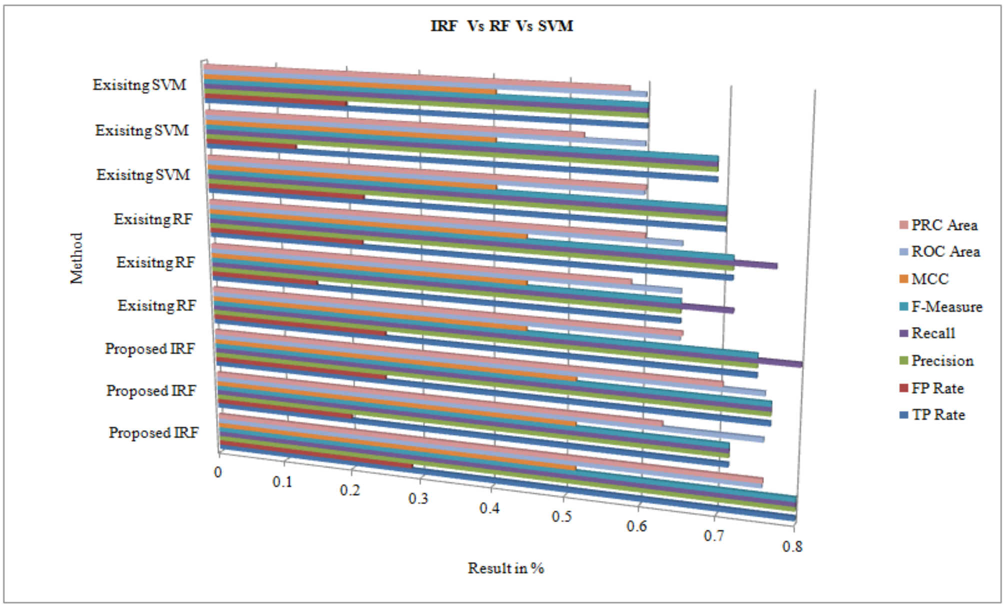

Figure 1.4 Simulation result for Proposed IRF Vs RF Vs SVM Method

Figure 1.4 shows the simulation graph for a total 30,200 Covid-19 dataset for various existing methods SVM, PR, and Proposed IRF. The simulation results clearly demonstrated that proposed IRF method performs outstanding in terms of TP %, FP %,Precision %, Recall %, F-Measure %, MCC %, ROC % and PRC % over existing PR and SVM methods.

5. Conclusion and Future work

Machine learning methods are widely used for data analytics in various fields. Currently the entire world is phasing the challenges of Covid-19 pandemic. Due to unavailability of proposer treatment, vaccines the Covid-19 infection are increasing rapidly in all over the world. This research paper presented an Improved Random Forest algorithm for predicting the COVID-19 pandemic patient health. This proposed method utilizes the quality of features of existing random forest method with some advanced features like ranking attributes and new voting features.

This proposed IRF method analyzes the available healthcare such as number of hospitals, ventilators, beds and treatments facility and similar Covid-19 patient patient's health data as normal and critical. This proposed IRF model helps in prediction of Covid-19 patient's health as well as in mapping of limited health resources with Covid-19 patient. The proposed IRF method and existing RF, SVM methods were implemented over WEKA Simulation tool and various performance measuring parameters TPR, FP, Recall, FP, TP, ROC, PRC and MCC were calculated. An experimental analysis clearly shows the proposed IRF method performs outstanding over existing methods. In future work the proposed IRF method will be tested on large live database of Covid-19 new datasets and compared with various existing methods.

References

[1] *Cohen, Joseph Paul, Paul Morrison, Lan Dao, Karsten Roth, Tim Q. Duong, and Marzyeh Ghassemi. "COVID-19 Image Data Collection: Prospective Predictions Are the Future." arXiv preprint arXiv: 2006.11988 (2020).*

[2] *Livingston, Edward, and Karen Bucher. "Corona-virus disease 2019 (COVID-19) in Italy." Jama 323, no. 14 (2020): 1335-1335.*

[3] *Bai, Yan, Lingsheng Yao, Tao Wei, Fei Tian, Dong-Yan Jin, Lijuan Chen, and Meiyun Wang. "Presumed asymptomatic carrier transmission of COVID-19." Jama 323, no. 14 (2020): 1406-1407.*

[4] *Metsky, Hayden C., Catherine A. Freije, Tinna-Solveig F. Kosoko-Thoroddsen, Pardis C. Sabeti, and Cameron Myhrvold. "CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design." BioRxiv (2020).*

[5] *Liu, Dianbo, Leonardo Clemente, Canelle Poirier, Xiyu Ding, Matteo Chinazzi, Jessica T. Davis, Alessandro Vespignani, and Mauricio Santillana. "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models." arXiv preprint arXiv:2004.04019 (2020).*

[6] *Ienca, Marcello, and Effy Vayena. "On the responsible use of digital data to tackle the COVID-19 pandemic." Nature medicine 26, no. 4 (2020): 463-464.*

[7] *Wynants, L., Van Calster, B., Bonten, M.M., Collins, G.S., Debray, T.P., De Vos, M., Haller, M.C., Heinze, G., Moons, K.G., Riley, R.D. and Schuit, E., 2020. Prediction models for diagnosis and prognosis of COVID-19 infection: a systematic review and critical appraisal. BMJ, 369.*

[8] *Ong, Edison, Mei U. Wong, Anthony Huffman, and Yongqun He. "COVID-19 corona-virus vaccine design using reverse vaccinology and machine learning." BioRxiv (2020).*

[9] *Li, Lin, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, et al. "Artificial intelligence distinguishes COVID-19 from community-acquired pneumonia on chest CT." Radiology (2020): 200905.*

[10] *Alimadadi, Ahmad, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe, and Xi Cheng. "Artificial intelligence and machine learning to fight COVID-19." (2020): 200-202.*

[11] *Randhawa, G.S., Soltysiak, M.P., El Roz, H., de Souza, C.P., Hill, K.A., and Kari, L., 2020. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: a COVID-19 case study. PloS one, 15(4), p.e0232391.*

[12] *Yan, Li, Hai-Tao Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li et al. "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan." MedRxiv (2020).*

[13] *Rao, Arni SR Srinivasa, and Jose A. Vazquez. "Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine." Infection Control & Hospital Epidemiology 41, no. 7 (2020): 826-830.*

[14] *Gates, B., 2020. Responding to Covid-19—a once-in-a-century pandemic?. New England Journal of Medicine, 382(18), pp.1677-1679.*

[15] *Tárnok, Attila. "Machine Learning, COVID‐19 (2019‐nCoV), and multi‐OMICS." Cytometry 97, no. 3 (2020): 215.*

[16] *Patil, Vipul, and Umesh Kumar Lilhore. "A survey of different data mining & machine learning methods for credit card fraud detection." International Journal of Scientific Research in Computer Science, Engineering and Information Technology 3, no. 5 (2018): 320-325.*

[17] *Trivedi, Naresh Kumar, Sarita Simaiya, Umesh Kumar Lilhore, and Sanjeev Kumar Sharma. "An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods."International Journal of Advanced Science and Technology, Vol. 29, No. 5, (2020), pp. 3414 - 3424.*

[18] *Inciardi, Riccardo M., Laura Lupi, Gregorio Zaccone, Leonardo Italia, Michela Raffo, Daniela Tomasoni, Dario S. Cani et al. "Cardiac involvement in a patient with corona-virus disease 2019 (COVID-19)." JAMA cardiology (2020).*

[19] *Vaishya, Raju, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem. "Artificial Intelligence (AI) applications for COVID-19 pandemic." Diabetes & Metabolic Syndrome: Clinical Research & Reviews (2020).*

[20] *Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R News 2, no. 3 (2002): 18-22.*

[21] *Bullock, Joseph, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. "Mapping the landscape of artificial intelligence applications against COVID-19." arXiv preprint arXiv: 2003.11336 (2020).*

[22] *Yan, Li, Hai-Tao Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li et al. "Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan." medRxiv (2020).*

[23] *Bonow, R.O., Fonarow, G.C., O'Gara, P.T. and Yancy, C.W., 2020. Association of corona-virus disease 2019 (COVID-19) with myocardial injury and mortality. JAMA cardiology.*

[24] *Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), pp.1947-1958.*

[25] *Chandra, Prakash, Umesh Kumar Lilhore, and Nitin Agrawal. "Network intrusion detection system based on modified random forest classifiers for kdd cup-99 and nsl-kdd dataset." International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 08 | Aug -2017, pp 786-791.*