Prediction of COVID-19 Using Machine Learning Classification Algorithms

Sherwin Vishesh Jathanna

Student, KMWA PU College, Bangalore, Karnataka, India

Abstract - *COVID-19, Corona Virus Disease-2019, caused by a novel Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2). An effective screening of this virus can enable quick and efficient diagnosis of COVID-19 can reduce the burden on the healthcare system. A detailed analysis on the provided dataset can build different and various types of machine learning algorithms, which their performance could be computed and further evaluated. In the following case Random Forest outperformed all the other Machine Learning models like SVM, Decision Tree, KNN & Logistic Regression.*

Key Words: SARS-Cov-2, Machine Learning (ML), SVM, KNN, Decision Tree, Random Forest, Logistic Regression

1. INTRODUCTION

The Corona Virus Disease 2019 (COVID-19), caused by SARS-CoV-2 infection [1-4], has spread around the world and became a global pandemic declared by World Health Organization on March 11, 2020[5, 6]. As of April 2021, the overall number of victims confirmed to carry this disease has reached up to 150 million in 219 countries and territories and almost 3 million deaths caused by this virus [7].The pandemic still continues to challenge the medical system all around the world and this raised a sudden demand for medical equipment, while the whole nation was in lockdown, the medical sector was heavily challenged by this virus and many brave health workers had to lose their lives. Currently, the validated diagnostic test for SARS-Cov-2 is by using reverse transcriptase-polymerase chain reaction (RT-PCR), and this has been in-shortage in developing and underdeveloped countries, this will lead in increase in cases and it may delay critical preventive measures to those who are in need. The effective diagnosis of COVID-19 can reduce the burden on the medical/healthcare system. Some prediction models combine several features to estimate like computer tomography (CT) scans [8-12], clinical symptoms [13], laboratory tests [14, 15], and integration of these given features [16]. In this paper, we put forward a machine-learning model that predicts whether a person is a carrier of SARS-CoV-2 infection by asking few basic questions. This model was trained by the data provided by kaggle.com [28]. Hence, this model can be implemented globally for effective screening and prioritization of testing for the virus for the general population.

2. LITERATURE SURVEY

SVM gives a great output while in comparison to different strategies for a great disorder prediction within algorithms [26]. This is a Machine Learning (ML) model that used a classification algorithm for two-group classification problems. What it does is basically builds a learning model that assigns new examples to one group or another. Statistic mining is useful for getting significant information. The data was collected from kaggle.com [28]

In some cases, ML was used in Bibliometric analysis of Covid-19 [18], classifying and detection of Covid-19 using X-Ray Images [19] [20], and surveillance of the disease by using genomically -comprehensive ML design[21] and some prediction models combine several features to estimate like computer tomography (CT) scans [8-12], clinical symptoms [13], laboratory tests [14, 15], and an integration of these given features [16] and further in symptoms there are papers that have used gradient-boosting machine model, LightGBM, auROC, SHAP which provides an precise output[19]

And also ML is been used in prediction for other diseases like Diabetes, Heart diseases, etc. [17-18].

3. METHODOLOGY AND MATERIALS USED

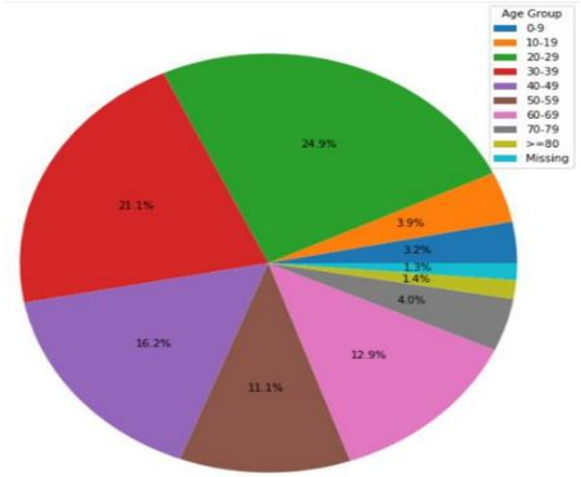
3.1 Materials Used

- The statistical data of 316831 cases were used in this study, which was collected from kaggle.com [28].
- Male, Female and Trans-gender sex were included in the study
- The age group were categorized into 1-categorizes and its represented in Figure 1, each having a difference of 9 years, and a missing(data not provided by the user) category
- The symptoms of COVID-19 are flu-like symptoms and its details are tabulated in table 1

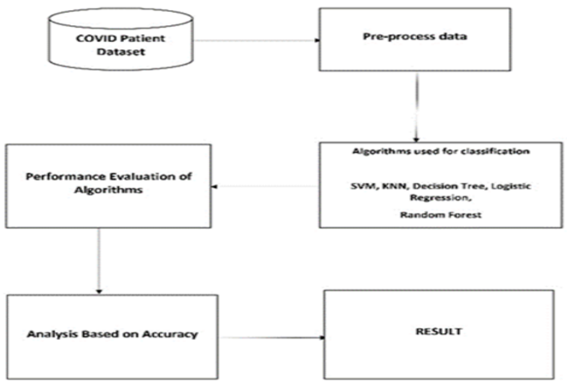
Table 1: *Symptoms of COVID-19*

Most Common Symptom	Moderate Symptoms	Severe Symptoms
Fever, Tiredness and Dry cough	Conjunctivitis, diarrhea, headache, aches, severe pains, sore throat, loss of the ability to taste or smell, skin rashes and chills	Shortness of breath, chest pain and loss of speech

Figure 1: *Percentage of COVID-19 cases per age group*



3.2 Flowchart


COVID Patient Dataset
Pre-process data
Algorithms used for classification
SVM, KNN, Decision Tree, Logistic Regression,
Random Forest
Performance Evaluation of Algorithms
Analysis Based on Accuracy
RESULT

3.3 ALGORITHMS USED

- **SVM**
- **KNN**
- **Decision Tree**
- **Random Forest**
- **Logistic regression**

3.3.1 SVM

SVM stands for Support Vector Machine, it is a standard supervised learning algorithm and this was introduced by Vapnik in 1995. This is a machine learning model that used a classification algorithm for two-group classification problems. What it does is basically builds a learning model that assigns new examples to one group or another. This model can be used for all the classification and regression challenges. The model makes a hyperplane and divides the data into classes resulting in all samples belonging to one particular class which could be categorized on one side and remaining on the other side. So we need to select the class which has a higher Margin (distance between the planes) [26].

3.3.2 KNN

KNN stands for K-Nearest Neighbor, this algorithm is the best method for classification which is based on similarities to other provided cases. The so-called others are nothing but called Neighbor. When a new case is provided, the distance from each case in the model is been calculated. Further applying this classification specifies the case as being the nearest neighbor, which may be highly similar. Furthermore, it puts the case into the group which contains the nearest neighbors. The algorithm is further able to calculate values continuously for that particular target. Hence, the average or (in some cases) median target value of the nearest neighbor is used to obtain or predict the value of the new case. [27]

3.3.3 Decision Tree

Decision tree is a supervised studying algorithm that consists of a set of rules. This algorithm is used frequently for the type of problems. In this particular algorithm, the facts as a whole are represented inside the shape of a tree wherein which each and every leaf is corresponding to a class label and the attributes correspond to the inner node of the tree. The fundamental opine is to discover the fundamental foundation node in each stage.

3.3.4 Random Forest

Random forest or Random Decision Forest is a supervised system, which is used for classification, regression and other tasks. In this, the algorithm consists of tree structures where the number of tree structures is directly proportional to the accuracy of the output or result, and where each internal node within that particular tree corresponds to an attribute and each and every leaf node represents a class label.

3.3.5 Logistic Regression

Logistic Regression is an algorithm that uses a logistic function which further models a binary dependent variable, many more complex extensions exists. The result given by this model is totally based on the opportunity feature. This uses a fee function which is referred to as sigma, and this function is more complex than the normal linear function. Logistic regression has limited the cost function which has a value between 0 and 1.

4. DATA SET

The statistic set is taken directly from Kaggle [28]. The statistics set has many attributes like Age, Symptoms, any contact with a carrier (Covid), etc. Records set is trained to get the accurate end result and further it is tested.

The accuracy is measured by the formula given by C5.0 algorithm which is,

ACCURACY = (TP+TN) / (TP+FP+TN+FN) Where the variables are, TP: True Positive TN: True Negative FP: False Positive FN: False Negative

5. RESULTS

The Result of the Machine Learning Models are as follows:

M.L. Models	Accuracy (In %)	Misclassification (In %)
SVM	86.667	13.333
KNN	83.333	16.667
Decision Tree	83.447	16.553
Random Forest	86.687	13.313
Logistic Regression	83.343	16.657

Table 02: *Result Description*

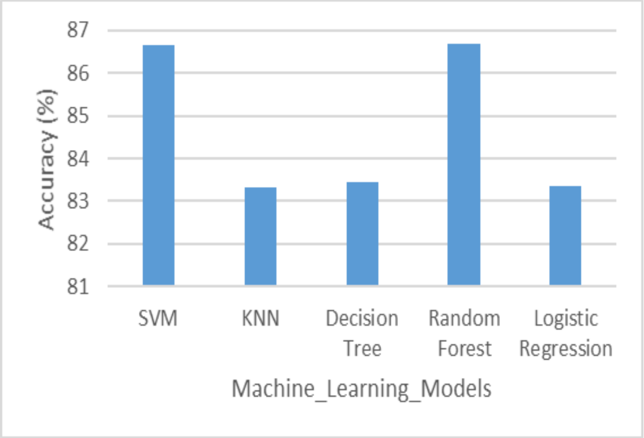


Figure 02: Bar Graph on *Result Description*

6. CONCLUSION

By using these five Machine Learning algorithms we have measured different parameters within the dataset and further Random Forest outperformed with an accuracy rate of approximately 87%. In future, more Machine Learning classifiers are required for evolving COVID-19 dataset.

7. ACKNOWLEDGEMENTS:

The author gratefully acknowledge Mr. Dinesh Acharya, Mr. Krishnamoorthy and Mr. Prakash Kalingrao Aithal, from Dept. of CSE, Manipal Institute of Technology(MIT), for guiding me in carrying out this study. I also thank Dr. Wilma Delphine Silvia CR, Professor and HOD, Department of Biochemistry, Sri Atal Bihari Vajpayee Medical College and Research Institute, Bangalore for reviewing and editing the manuscript.

8. REFERENCES

[1] Lamers MM, Beumer J, van der Vaart J. et al. SARS-CoV-2 productively infects human gut enterocytes. Science. 2020;369:50-54

[2] Wang QH, Zhang YF, Wu LL, Niu S. et al. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell. 2020;181:894-904

[3] Yan RH, Zhang YY, Li YN. et al. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science. 2020;367:1444-1448

[4] Shang J, Ye G, Shi K. et al. Structural basis of receptor recognition by SARS-CoV-2. 2020;581:221-224.

[5] Sica A, Vitiello P, Caccavale S. et al. Primary cutaneous DLBCL non-GCB type: challenges of a rare case. Open Med (Wars). 2020;19:119-125

[6] World Health Organization. Coronavirus disease 2019 (COVID-19) situation report 2020. https://covid19.who.int/

[7] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. https://doi.org/10.1016/S1473-3099(20) 30120-1 (2020).

[8] Gozes, O. et al. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. arXiv e-prints 2003, arXiv:2003.05037 (2020).

[9] Song, Y. et al. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. medRxiv https://doi.org/10.1101/2020.02.23.20026930 (2020).

[10] Wang, S. et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv, https://doi.org/10.1101/2020.02.14.20023028 (2020).

[11] Jin, C. et al. Development and evaluation of an AI system for COVID-19 diagnosis. medRxiv, https://doi.org/10.1101/2020.03.20.20039834 (2020).

[12] Punn, N. S. & Agarwal, S. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. arXiv:2004.11676 [cs, eess] (2020).

[13] Tostmann, A. et al. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. Eurosurveillance 25, 2000508 (2020).

[14] Feng, C. et al. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. medRxiv, https://doi.org/10.1101/2020.03.19.20039099 (2020).

[15] Punn, N. S., Sonbhadra, S. K. & Agarwal, S. COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. medRxiv, https://doi.org/10.1101/2020.04.08.20057679 (2020).

[16] Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat. Med. 26, 1224–1228 (2020).

[17] Prediction Of Diabetes Using Machine Learning Classification Algorithms, 2277-8616

[18] Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases, 2088-8708

[19] Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods

[20] Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach, arXiv:2004.10641v1 [eess.IV] 22 Apr 2020

[21] CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design, https://doi.org/10.1101/2020.02.26.967026;

[22] Kolla, B.P. & Raman, A.R. 2019, Data Engineered Content Extraction Studies for Indian websites, Advances in Intelligent Systems and Computing, 711, pp. 505-512.

[23] Prakash, K.B. 2017, "Content extraction studies using total distance algorithm", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 673.

[24] Prakash, K.B., Kumar, K.S. & Rao, S.U.M. 2017, "Content extraction issues in online web education", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. https://doi.org/10.1109/ICATCCT.2016.7912086

[25] Prakash, K.B., Rajaraman, A., Perumal, T. & Kolla, P. 2016, "Foundations to frontiers of massive data analytics", Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016, pp. 242.

[26] Vapnik, V. N. The nature of statistical learning theory. New York: Springer, 1995.

[27]. Yazdani A, Ebrahimi T, Hoffmann U. Classification of EEG signals using Dempster Shafer theory and a K-nearest neighbor classifier. IEEE. In: Proc of the 4th int EMBS conf on neural engineering, 2009: 327-30.

[28] https://www.kaggle.com/iamhungundji/covid19-symptoms-checkerc