Severity Prediction of COVID-19 Patients Using Machine Learning Classification Algorithms: A Case Study of Small City in Pakistan with Minimal Health Facility

Hina Gull
Department of Computer Information Systems, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box No. 1982, Dammam, Saudi Arabia
e-mail: hgull@iau.edu.sa

Gomathi Krishna*
Department of Computer Information Systems, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box No. 1982, Dammam, Saudi Arabia
e-mail: gkrishna@iau.edu.sa

May Issa Aldossary
Department of Computer Information Systems, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box No. 1982, Dammam, Saudi Arabia
e-mail: mialdossary@iau.edu.sa

Sardar Zafar Iqbal
Department of Computer Information Systems, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box No. 1982, Dammam, Saudi Arabia
e-mail: saiqbal@iau.edu.sa

2020 IEEE 6th International Conference on Computer and Communications (ICCC) | 978-1-7281-8635-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCC51575.2020.9344984

Abstract—**Coronavirus disease has been declared as an infectious pandemic affecting the life and health of millions across the globe. It has caused high number of mortalities giving birth to exceptional state of emergency worldwide. It has not affected the people but also has damaged infrastructure of different countries, especially causing an expectational situation in health care systems globally. Due to unavailability of vaccination and faster human to human transmission of virus, healthcare facilities are at high risk of exceeding their limit and capacity, especially in developing countries like Pakistan. Therefore, it is important to manage resources properly in these countries to control high mortality rate and damage it can cause. In this paper we have taken a case study of small city in Pakistan, where healthcare facilities are not enough to deal with pandemic. Most of the COVID-19 patients have to be refer to big cities based on their severity. We have taken data of COVID-19 positive patients from this small city, developed and applied machine learning classification model to predict the severity of patient, in order to deal with the shortage of resources. Among all seven taken and tested algorithms, we have chosen SVM to predict severity of patients. Model has shown 60% of accuracy and have divided patient's severity into mild, moderate and severe levels.**

Keywords-severity level; Machine learning Algorithms; limited resources; healthcare facilities

I. INTRODUCTION

Coronavirus Disease named as COVID-19 is a speedily emergent respiratory track disease [1] which is caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). Virus was [2] first emerged in Wuhan China in late 2019 and rapidly become pandemic causing huge number of infected cases and high mortality rates. As of today, July 14, 2020, 216 countries with 12 929 306 confirmed cases are reported by World Health Organization due to faster human to human transmission. Asymptomatic patients have also become a vital source of infection dispersal. This pandemic has caused many countries to put people in lockdown, seal their borders and shutdown major economic activities, eventually effecting normal life of people to major extent.

Pakistan is also one of the countries majorly hit by COVID-19 with 253,604 confirmed cases of COVID-19 [3]. For the developing country like Pakistan, this pandemic is a high-risk trial for the public health and health care systems. This is because Pakistan has limited number of human and material resources related to health care. Especially in small cities and towns, lack of resources is a major hinderance in dealing with critical patients and control the mortality rates. With very limited number of resources and poor healthcare infrastructure, Pakistan has satisfactorily raised the level of readiness to combat COVID-19[4]. However, there is still needed to manage healthcare resources properly in small towns and cities to effectively deal with the Pandemic. Muzaffarabad is a small city in Pakistan with very poor healthcare infrastructure. City has two major hospitals with very limited resources to provide health care facility to local people. Government decided to urgently convert one of the governmental buildings into isolation hospital in March 2020 when COVID-19 hit this small town with limited beds and other resources such as ventilators. As number of positive cases were increasing in city, it was very difficult for hospital authorities to allocate beds and other facilities to critical patients and effectively manage the resources ahead of time with respect to the severity of patient's condition. This paper has taken data of COVID-19 positive patients of this small city of Pakistan, in order to help authorities in combating with pandemic by predicting the severity of COVID-19 positive patients. We have applied different machine learning algorithms to predict severity level of patient. It will help authorities to not only decide the clinical route to the recognition of critical cases, but also help them to manage their resources according to the priority of patient's health.

II. Background Study

In this section we have reviewed some of the research works similar to our work. We have analyzed the literature based on samples used, number of samples taken for analysis, limitations encountered during the study period.

Jiangpeng Wu et al.[5], proposed study emphasizes prompt treatment by earliest diagnosis. The researcher's study on 11 key blood indexes were mined using a random forest algorithm. The commercial blood test equipment is used to derive 49 clinical blood test data which is used to build the final tool for assistant discrimination. The main challenge faced in this study was, the requirement for the detection of COVID-19 cases with a standard symptom is not verified as these cases have been difficult to obtain in the present situation.

Qi, Xiaolong, et al. [6] to forecast hospital stays in patients with pneumonia affected with Covid-19, machine learning based model was built using logistic regression and random forest algorithms. The data is resourced from 52 infected patients with Covid-19 and their CT images collected from 5 selected hospitals in 5 different cities in China between Jan 23, 2020 and Feb 8, 2020. The proposed models accurately calculating hospital stay in patients with pneumonia Covid-19 infection. The hospital stay result is either minimum hospital stays with less than 10 days or maximum hospital stay that extends more than 10 days.

Rahul Kumar, et al. [7], Covid-19 virus spread is forecast using chest X rays of asymptomatic patients. This kind of prediction would be useful in spotting the epidemic early, which in turn will help to monitor it. A count of 5840 X-ray images from Italian public dataset are used for building a model and for final classification by machine learning techniques. This research classifies the public in to three categories namely COVID-19, Normal, and Pneumonia peoples.

Rodolfo M.Pereira et al. [8], recommended Chest X-ray (CXR) since it is cheaper, quicker and more commonly used. This study distinguishes COVID-19-induced pneumonia from other types using X-Ray images. The seven diverse classes of 1144 X-ray images are taken for this research purpose. The study does not include a conclusive COVID-19 diagnosis, but it helping to screen patients in Emergency services.

X. Xu et al. [9], for the primary detection of COVID-19 patients, the deep learning model developed by using pulmonary CT images, and this research could be a accompanying analytical tool for the Medics. This model identifies COVID-19 images from healthy images. In some cases, it cannot distinguish other kind of pneumonia with the COVID-19. The main drawback of this study was limited number of samples, and it's not adequate to train and test the model with better accuracy. To get better prediction additional details could be inquired namely contact round of patient, history of visiting places, early symptoms, and test center results.

Davide Brinati et al [10], The likelihood of using plasma test and machine learning to identify COVID-19 positive cases has been demonstrated as a substitute to nasopharyngeal swab tests. This research resourced from 279 patients' blood test data admitted at Hospital in Italy. In developing countries grief from scarcities of rRT-PCR substances and dedicated laboratories, it is very beneficial to consider blood testing to classify COVID-19 positive patients. Some of the limitations in this study are discussed by the authors. The comparatively limited number of cases considered is the first constraint. In order to control any unfairness, this was done by conducting nested cross-validation and by employing models that are proven to be successful with discreetly sized samples. Second, the study lacks generalization because it was a single-center study.

João Matos et al, [11], focused to measure disease volume (VoD) using a quick, easy and readily available post-processing CT method. Along with CT and clinical data predicts, patient's survival rate as early as possible. This analysis helps to decide and allocate ventilation facilities. Some of the limitations in this study are susceptible to selection bias among patients, and patients with very slight symptoms are excepted from the study. And also, this study lacks generalization because it was a single-center study.

III. Problem Statement

For this study, we would like to predict the severity of COVID-19 positive patients in order to analyze the health condition of patient in one of the small cities (Muzaffarabad) of Pakistan. Muzaffarabad is small city in Pakistan, where health facility is not good enough with very limited resources. Normally patients with severe conditions are shifted to nearby big cities in order to combat with patient's criticality. As COVID-19 started emerging in city, authorities have established a very small quarantine center with limited beds and other medical resources. As number of positive cases are increasing in city due to second wave of pandemic, it is very difficult for hospital authorities to allocate beds and other facilities to critical patients. Our machine learning model will predict the severity level of COVID-19 patient in order to help authorities to prioritize patients in order to allocate resources or to shift them to nearby city. Health condition prediction will not only help authorities to manage their resources but also help patients to get medical care as soon as possible based on their condition.

IV. Experimentation

Real open source data of COVID-19 patients with symptoms [12] (downloaded from Kaggle) was used to develop a prediction model. The dataset contains 992 records with symptoms like Fever, Tiredness, Dry-Cough, Difficulty-in-Breathing, Sore-Throat, Pains, Nasal-Congestion, Runny-Nose', 'Diarrhea', Age and Gender are used as attributes while condition (Mild, Moderate and Severe) is used as target variable. Experiments were conducted using python in PyCharm. Data set is divided into 80% of training data and 20% of testing data. In the training set, the model is built from the classification techniques. In order to check which of the algorithm will best suit in prediction, we have tested following seven classification algorithms:

- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN).
- Classification and Regression Trees (CART).
- Gaussian Naive Bayes (NB).
- Support Vector Machines (SVM).
- Random Forest Selection (RF)

To select the best model for prediction, following measures were used to evaluate the classification quality:

A. *Accuracy*

Given a dataset consisting of (TP + TN) data points, the accuracy is equal to the ratio of total correctly prediction items (TP + TN + FP + FN) and the total data points in the sample [13].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Figure 1 and Table I shows comparison of machine learning algorithms in terms of accuracy.

TABLE I. ACCURACY OF ALGORITHMS

Algorithm	Accuracy
LR	0.547168
LDA	0.566123
KNN	0.500459
CART	0.500506
NB	0.508085
SVM	0.602706
RF	0.510665

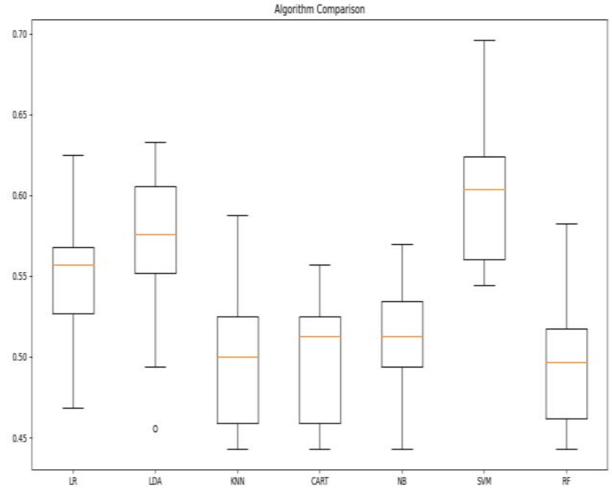

Figure 1. Accuracy of Algorithms (Comparision).

B. *Precision*

Precision is considered to be the ratio of Positive samples that are true (TP) and the total of both Positively True and Positively False samples (FP) [13]

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

C. *F1 Score*

F1 Score is equal to the harmonic mean of Recall and Precision value [14].

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (3)$$

TABLE II. UNITS VS CALCULATED VALUES

Unit	Value
Precision	0.95
Recall	0.69
F-Score	0.80

V. RESULTS

Real data of the patients tested positive for COVID-19 in Muzaffarabad is taken. All their symptoms are recorded and SVM is applied on collected data to predict the severity level of patients as shown in Table III. Data is shown in table III and explained as follows: Fever (represented by F) =1 or 0, Tiredness (represented by T) = 1 or 0, Dry Cough (represented by DC)= 1 or 0, Breathing Problem (represented by BP)=1 or 0, Sore Throat (represented by ST)= 1 or 0, Pains(represented by P)= 1 or 0, Nasal Congestion((represented by NC)= 1 or 0, (represented by RN)= 1 or 0, Diarrhea(represented by D)= 1 or 0, Age (represented by A): Age 0-9=1, Age 10-19=2, Age 20-24=3,Age 25-59=4,Age 60+=5, Gender (represented by G): Male=1, Female=2.

Based on symptoms observed by the patients, SVM has predicted severity level of patients, which will help authorities to deal with lack of resources, decide a clinical route for severely critical patients by prioritizing the patients according to their condition. It has been observed from the table below most of the people having all symptoms and age between 25-59 are more at risk. It can also be observed from the table that people who are more at severe risk are women of age between 25-59. It has also been observed from the data that most of the patients in city are of age 25-59 which are affected by the virus. Data also depicts the fact that patients having all symptoms are severely affected by the virus. To be specific about the considered attributes of COVID-19, it was observed that very few people are having breathing problem, runny nose and diarrhea, while fever, tiredness and dry cough are moderately observed in people suffering from COVID-19.

TABLE III. PREDICTION

F	T	DC	BP	ST	P	NC	RN	D	A	G	Prediction
1	1	1	0	0	0	0	0	0	4	1	Mild
0	0	0	0	0	0	0	0	0	4	1	Mild
0	0	0	0	0	1	1	0	0	4	1	Mild
0	0	0	0	0	0	0	0	0	4	2	Mild
1	1	1	0	1	0	0	0	0	4	1	Mild
1	1	1	1	1	1	1	0	0	4	2	Severe
0	1	1	0	0	0	1	1	0	4	2	Moderate
1	1	1	1	1	0	1	0	0	4	2	Severe
0	0	0	0	0	0	0	0	0	4	2	Mild
1	1	1	1	1	0	1	0	0	3	1	Moderate
1	0	1	0	1	1	1	0	0	4	2	Moderate
1	1	1	0	0	0	0	1	0	4	1	Mild
1	0	0	1	1	1	0	0	0	4	2	Moderate
0	0	0	0	0	0	0	0	0	5	1	Mild
1	0	0	0	0	0	0	0	0	3	2	Mild
1	0	0	0	0	0	0	0	0	3	1	Mild
1	0	0	0	0	0	0	0	0	4	2	Mild
1	1	0	0	1	1	1	0	0	5	1	Moderate
1	0	0	0	1	0	0	0	0	4	1	Mild
1	1	1	1	1	1	1	0	0	4	2	Severe
1	1	1	1	1	1	1	0	0	4	1	Severe
1	1	1	1	1	1	1	0	0	5	2	Severe
1	1	1	1	0	0	0	0	0	4	1	Moderate
1	0	0	0	0	1	0	0	0	4	1	Moderate
1	1	1	1	1	1	0	0	0	3	2	Severe
1	1	1	1	1	1	1	0	0	3	2	Severe
1	1	1	1	0	0	0	0	0	2	1	Mild
1	1	1	1	1	1	0	0	0	3	2	Severe
1	1	1	1	1	1	1	0	0	4	2	Severe
1	1	0	0	1	1	1	0	0	5	1	Moderate
1	0	0	0	1	0	0	0	0	4	1	Mild
1	0	1	0	1	1	1	0	0	4	2	Moderate
1	1	1	1	1	0	1	0	0	4	2	Severe
1	1	1	0	0	0	0	1	0	4	1	Mild
1	1	1	1	1	1	1	0	1	4	2	Severe
1	1	1	1	1	1	0	1	1	4	2	Severe
1	1	1	0	1	0	0	1	0	4	1	Mild

Results obtained can also be depicted form the graph below. Figure 2 shows the symptoms in people which are predicted as Mild. As depicted in the graph, people predicted with mild severity shows varying symptoms. Fever is the most obvious symptom seen in most of the patients. 68.75% are suffering from fever, 37.5% are suffering from Tiredness and dry cough, 25% are having sore throat. Diarrhea was only symptoms which was not observed in any of the patients having mild severity level.

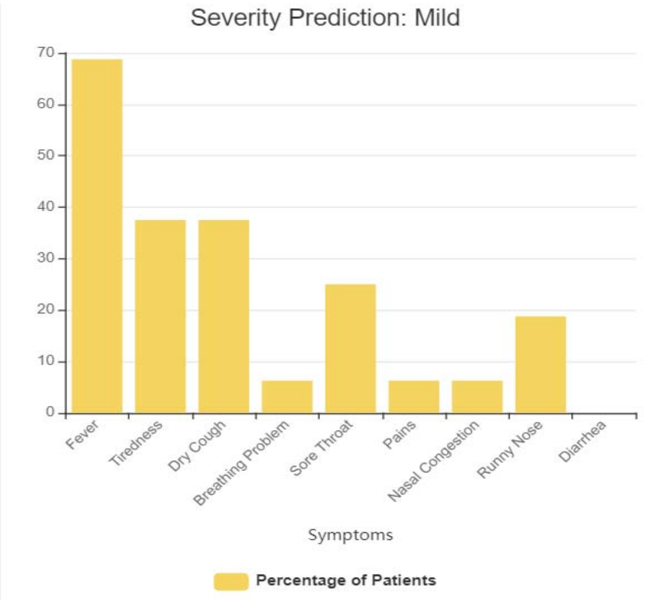

Figure 2. Severity Prediction: Mild v/s Patient %age

Figure 3 shows the symptoms in people who are predicted as Moderate. As depicted in the graph, people predicted with moderate severity shows varying symptoms. Pains is the most obvious symptom seen in most of the patients. 52.94% are suffering from fever, 29.41 and 37.5% are suffering from Tiredness and dry cough respectively. Like Mild severity, Diarrhea was only symptoms which was not observed in any of the patients having mild severity level.

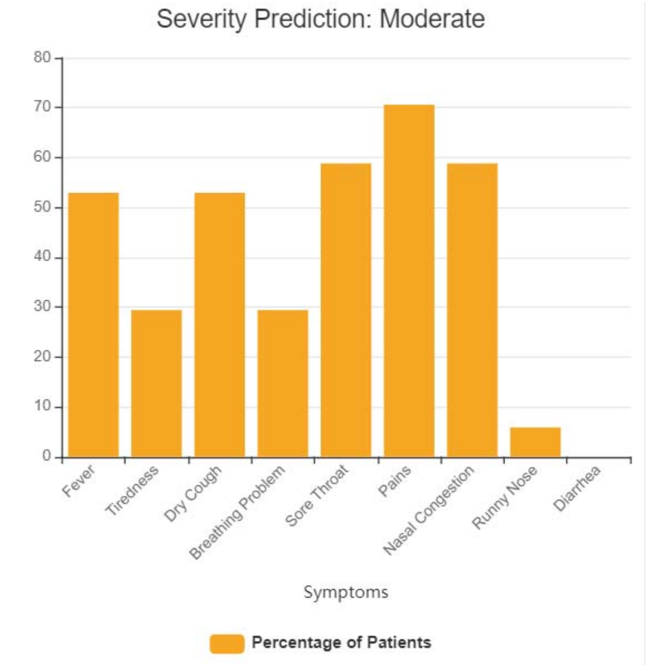

Figure 3. Severity Prediction: Moderate v/s Patient %age

Figure 4 shows the symptoms in people which are predicted as Severe. As depicted in the graph, people predicted with moderate severity shows almost all the symptoms. All patients are having fever, tiredness, dry cough, breathing problem and sore throat. Unlike mild and moderate severity, Diarrhea was also observed in many patients.

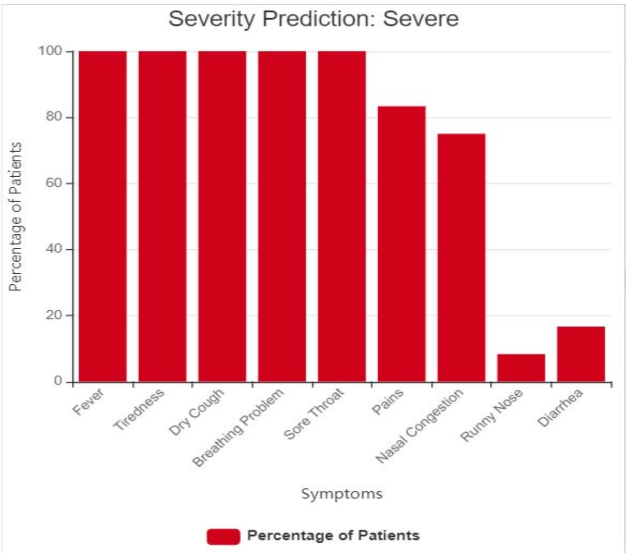

Figure 4. Severity Prediction: Severe v/s Patient %age

VI. Conclusion

The data resourced for this study, taken from one of the small cities of Pakistan. The dataset contains 992 records of COVID-19 affected patients and eleven different attributes are considered for this research. The different classification algorithms are applied on this data, among these Support Vector Machine (SVM) identified as better classification algorithm. This early detection of COVID patients, support concern higher authorities to take better decision and help the people by providing better services with limited available resources. And also, this helps to isolate the affected people as early as possible, and it prevents the spread of dangerous pandemic.

References

[1] S. H. A. Khoshnaw, M. Shahzad, M. Ali, and F. Sultan, “A quantitative and qualitative analysis of the COVID–19 pandemic model,” Chaos, Solitons and Fractals, 2020, doi: 10.1016/j.chaos.2020.109932.

[2] WHO, “Novel Coronavirus (2019-nCoV),” WHO Bull., 2020.

[3] A. Waris, U. K. Atta, M. Ali, A. Asmat, and A. Baset, “COVID-19 outbreak: current scenario of Pakistan,” New Microbes and New Infections. 2020, doi: 10.1016/j.nmni.2020.100681.

[4] N. Noreen et al., “Coronavirus disease (COVID-19) Pandemic and Pakistan; Limitations and Gaps,” Limitations Gaps. Glob. Biosecurity, 2020.

[5] J. Wu et al., “Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results,” 2020, doi: 10.1101/2020.04.02.20051136.

[6] H. Yue et al., “Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study,” Ann. Transl. Med., 2020, doi: 10.21037/atm-20-3026.

[7] R. Kumar et al., “Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers,” pp. 1–10, 2020, doi: 10.1101/2020.04.13.20063461.

[8] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla, and Y. M. G. Costa, “COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios,” Comput. Methods Programs Biomed., 2020, doi: 10.1016/j.cmpb.2020.105532.

[9] X. Xu et al., “Deep learning system to screen coronavirus disease 2019 pneumonia,” arXiv. 2020.

[10] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, “Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study,” J. Med. Syst., 2020, doi: 10.1007/s10916-020-01597-4.

[11] J. Matos et al., “Evaluation of novel coronavirus disease (COVID-19) using quantitative lung CT and clinical data: prediction of short-term outcome,” Eur. Radiol. Exp., 2020, doi: 10.1186/s41747-020-00167-0.

[12] “COVID-19 Symptoms Checker-Kaggle,” 2020. https://www.kaggle.com/iamhungundji/covid19-symptoms-checker.

[13] M. Hofer, G. Strauß, K. Koulechov, and A. Dietz, “Definition of accuracy and precision-evaluating CAS-systems,” Int. Congr. Ser., 2005, doi: 10.1016/j.ics.2005.03.290.

[14] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, 2020, doi: 10.1186/s12864-019-6413-7.