COVID-19 Mortality Prediction using Machine Learning Methods

Akashdeep Singh Rana, M.Tech. Research Scholar
Department of CSE, SBBS University, Jalandhar,
(akash511969@gmail.com)

Dr.Harmeet Singh, Associate Professor,
Department of CSE
SBBS University, Jalandhar
(hsnatt5@gmail.com)

Dr.VijayDhir, Professor, Department of CSE,
SBBS University, Jalandhar
(drvijaydhir@gmail.com)

***Abstract*:**

COVID-19 pandemic affects the world disastrously and also had a major effect on the world economy. We aimed to create the prediction model of in-hospital mortality using machine learning methods for patients with coronavirus disease 2019 (COVID-19) based on Chronic disease, age, smoking, and gender. The model was applied to the reliable data published by the government of Mexico and the dataset was collected by Mexican health authorities. The important variables used in this model are age, hypertension, gender, COPD, smoking, intubated, diabetes, asthma, pneumonia, and cardiovascular disease. Furthermore, we calculated Accuracy using data from January 1, 2020, to April 26, 2022. Only those patients who had full information were included in this study. Of the 43,110 patients admitted for COVID-19, 3864 (8.91 %) died during their stay. Linear Regression, Decision Tree, Naive Bayes, and K-Nearest Neighbour have been used to build the model. The Model provided an excellent result with an accuracy of 89.34 %. The model can be useful in estimating the in-hospital mortality of COVID-19 patients and minimizing the deaths due to COVID-19.

Keywords*: *COVID-19, Prediction Model, Hybrid Model,Machine Learning.
DOI Number: 10.14704/nq.2022.20.8.NQ44230 **NeuroQuantology 2022 ;20(8):2113-2117**

1. INTRODUCTION

COVID-19 is an infectious disease caused by a coronavirus. It was discovered in December 2019 in Wuhan, China. COVID-19 has been classified as a pandemic by WHO due to an increase in cases around the world[1]. As of May 6, 2022, a total of 526,092,380 cases have been reported, with 519,813,714 successfully treated. Unfortunately, 6,278,666 people have died as a result of this virus. [2]. Moreover, it was found that patients with ages above 65 have a higher risk of in-hospital mortality. The elderly and those with comorbid conditions, particularly heart disease had the highest mortality rates[3]. Patients who died as a result of coronavirus had a higher likelihood of having COPD and being current smokers[4]. Self-quarantine and distancing are advised for those suffering from mild coronavirus symptoms. This disease affects each individual differently. People with a weak immune system are more likely to get this disease quickly than people with a strong immune system[5]. The spread of coronavirus is more likely as a result of failing to observe social distancing. This virus then spreads from person to person and continues to spread[6].

Many researchers are working to reduce the impact of this disease. Machine Learning (ML) has the ability to establish the correlation between deaths and disease spread. The WHO, world communities, and kaggle are all working to update patient data so that researchers can use it to stop or slow the spread of this virus. Because the data of coronavirus patients are so complex, establishing a relationship between them is extremely difficult. Many Machine Learning (ML) models have been developed to discover the complex relationship between coronavirus data. The ML primarily based totally version can assist to expect the COVID-19 pandemic spread, the wide variety of cases, deaths, and effective measures to prevent the pandemic.

Many papers employ the Machine learning technique to create a model for COVID-19. One paper proposed a model to predict the number of deaths due to the corona virus. In this work, Susceptible-Infected-Recovered (SIR) model was used to forecast the death of COVID-19 patients[7]. In this work it was found that lower the hemoglobin and lymphocytes lower the risk of death due to this virus. Another random forest model discovered that patients with diabetes mellitus have a higher risk of death from coronavirus[8].

The majority of the models were time-based. The goal of this model is to predict the death of an individual patient suffering from coronavirus based on chronic disease, age, gender, and smoking. Hypertension, COPD, intubation, diabetes, asthma, pneumonia, and cardiovascular disease are examples of chronic diseases. Age is also one of the most important inputs for predicting death.

Abbreviations

LR Linear Regression
KNN K-Nearest Neighbor
ML Machine Learning
NB Naive Bayes
DM Diabetes Mellitus
COVID-19 coronavirus disease-2019
COPD Chronic Obstructive Pulmonary Disease
DT Decision Tree

I. DATASET

The model was applied to a dataset collected by Mexican health authorities. The dataset includes patients who were hospitalized in Mexico, as well as their age, gender, date of death, and chronic diseases such as COPD, diabetes, asthma, hypertension, and cardiovascular disease. The dataset also includes the date of onset of symptoms, the date of admission, and the date of discharge/death.

II. MATERIAL AND METHODS

A Hybrid ML algorithm is used to predict the mortality risk prediction. Fig. 1, depicts the proposed hybrid model architecture with chronic diseases such as diabetes, COPD, hypertension, asthma, and cardiovascular disease. Gender, age, and smoking are also inputs. The output is the COVID-19 patient's death.

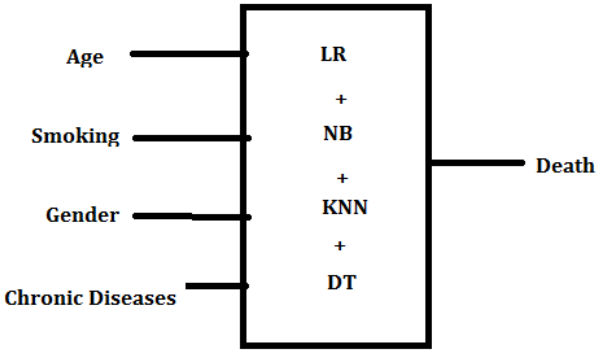

Fig. 1.: A general structure of a Hybrid model

The hybrid model has four models i.e., LR(Linear Regression), NB(Naive Bayes), KNN(K-Nearest Neighbor), and DT(Decision Tree) which will be ensemble using the voting classifier. In the input we give the patient's age, smoking tendencies, gender, intubated, and chronic diseases like COPD, diabetes, asthma, hypertension, Kidney, and cardiovascular disease.

A. Data Cleaning

A data cleaning approach is applied to remove the missing values. Patients with more than 50% missing values were removed. Like patients whose gender is unknown or having an outlier. Those whose value is unknown were removed from the dataset. As a result, the total dataset from 1046700 become 43,110.

*B. **Train Test Split***

After the cleaning of the dataset, the dataset is split into an 80:20 ratio. As a result, the number of samples of train and test becomes 34488 and 862. The dataset has 12 columns i.e., diabetes, COPD, asthma, hypertension, cardiovascular, obesity, kidney disease, smoking, intubated, pneumonia, age, and gender

*e*ISSN1303-5150

whereas in the output the dataset contains the target column which has the patient's death record. The training set of the target column was unbalanced that containing the number of deaths and the number of alive patients 3073, and 31415 respectively as shown in Fig. 2.

Training Set	Dead	3073
	Survived	31415
	Total	34488
Testing Set	Dead	791
	Survived	7831
	Total	8622

Fig. 2. Summary of the training set and test set of the dataset

C. *Data preprocessing*

a) Outlier Treatment

It is the first step of the preprocessing which is very useful to make a model efficient. In this step, extreme values were removed or the values other than numerical were also removed. This technique limits the extreme value of numerical data.

b) Feature values transformation

After the removal of outliers, the dataset becomes more efficient than before. As we know ML does not deal with units like kg, gram, inch, feet, etc. So we need to convert the data into the same format for making the data easy to understand for ML. It can be done using scaling like standard scaling, and min-max scaling. By using the scaling data is converted into a suitable numerical format for the Machine Learning technique.

c)Missing values imputation

After the scaling, the missing values can be replaced using KNN. In this algorithm, the missing value can be replaced using the mean value of the surrounding neighborhoods. This algorithm was compatible with the continuous feature as well as the categorical feature [9]. Each of the missing values in the dataset can be replaced using the KNN technique.

d) Oversampling

This is the final step of the preprocessing. As the issue of imbalance arises here. As shown in "Fig. 3", in the This is the final step of the preprocessing. As the issue of imbalance arises here. As shown in Fig. 3, in the training set the survivor was much more. The inequality case dominance toward the dominant class [10]. The number of dead patients in the training dataset was 3073 whereas survivors were 31415. Thus this makes the model more inefficient and creates an imbalance. To make the dataset balanced the oversampling technique can be used.

Training set	Dead	31415
	Survived	31415
	Total	62830

Fig. 3. Summary of the training set of the dataset After oversampling

The oversampling technique has the ability to increase the number of minority classes. Here the number of death is much lower than the number of survivors. Hence oversampling technique makes the number of death equal to the number of survivors. We used Synthetic Minority Oversampling Technique (SMOTE)algorithm for oversampling. The training set after the oversampling is shown in "Fig. 3". Only the training dataset was oversampled as test data used for the actual interpretation.

D. *Feature Selection*

After the oversampling of the dataset. The important features were extracted so that the model can be more efficient. As there were so many features that has the least importance to the model. It can be done using the recursive feature technique(RFE) around classifiers LR, gradient boosting, and AdaBoost. These algorithms have vast capabilities and can be applied in COVID-19 research[11][12].

Every feature coefficient metric was ranked according to the voter system. Features that had the lowest importance in the training set were removed using the RSE. The feature that was left behind was used for mortality prediction.

*e*ISSN1303-5150

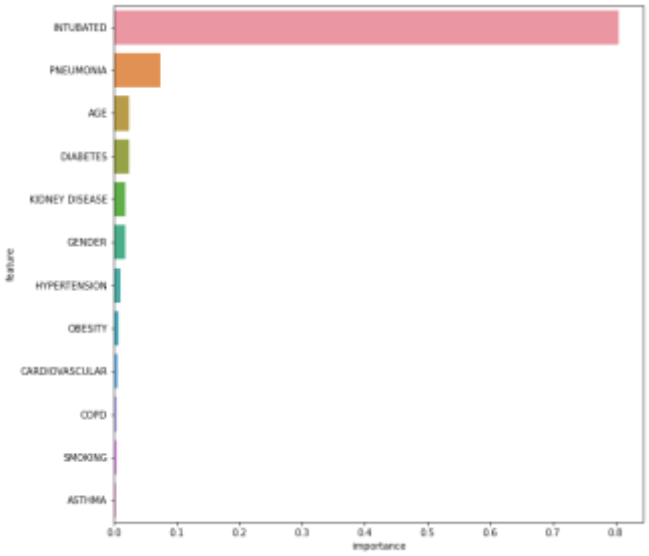

Fig. 4: Features with their importance in the model

As Fig. 4 shows that asthma had the lowest importance and intubated with the highest importance. Hence we removed the feature with the lowest importance from the dataset and continue with the remaining parameters.

E. Mortality Risk Prediction

After the feature selection, a hybrid model was created using NB, KNN, LR, and DT. To predict the mortality using the selected feature. This technique found useful to find the COVID-19 risk[13][14][15][16] .This model was an ensemble using the voting classification.

F. Results

It is possible that with some combinations of health problems or factors, the curve of probabilities is not the expected one, this may be because there are few patients who meet that combination of health problems or diseases, causing the prediction of the model to be somewhat doubtful. This problem seems to appear in those features or indicators in which the positive values are scarce. The latter can be consulted in the 'Final data display' section.

Based on the graphs created from the predictions of the model as shown in Fig. 5, we see that the disease with the highest risk of death is diabetes followed by COPD. However, we see that these curves grow when the patient has pneumonia, but this curve grows aggressively when patients have been intubated and diagnosed with pneumonia.

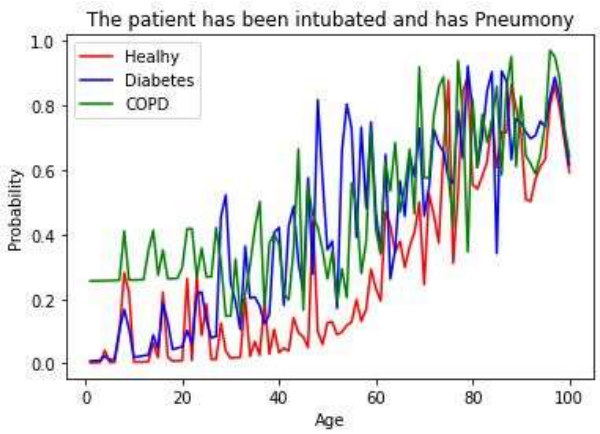

Fig. 5: Patients with intubated and pneumonia risk analysis

Based on the graphs created from the predictions of the model, we see that the disease with the highest risk of death is diabetes followed by COPD. However, we see that these curves grow when the patient has pneumonia, but this curve grows aggressively when patients have been intubated and diagnosed with pneumonia.

Evaluation Metric in %	
Accuracy	89.34
Precision	70
Recall	77.5
F1 score	73

Fig. 6: The evaluation result of the mortality prediction models

In Fig. 6 the performance of the model is shown which shows the accuracy, precision, recall, and f1 score.

IV. CONCLUSION

In this research, we have proposed a hybrid-based model to predict the mortality of COVID-19 patients. Age, COPD, and diabetes had a very high impact on the mortality of COVID-19 patients. The dataset is taken from the Mexican health authorities, which is publicly accessible to create models. As it was found that the feature asthma had the least impact on the model so it was removed. Hence to make the model more efficient. Using SMOTE algorithm the imbalanced feature in the dataset i.e., a target that includes a number of deaths and survivors was balanced. So that the majority class could not overcome the minority and make the model much more efficient when applied in the real world.

*e*ISSN1303-5150

REFERENCES

[1]Jarndal, Anwar, et al. "GPR and ANN based prediction models for COVID-19 death cases." 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). IEEE, 2020.

[2] Johns Hopkins University CORONAVIRUS RESOURCE CENTER, (https://coronavirus.jhu.edu/).

[3] Phillips, Matthew C., et al. "Effect of mortality from COVID-19 on inpatient outcomes." Journal of medical virology 94.1 (2022): 318-326.

[4]Yadaw, Arjun S., et al. "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model." The Lancet Digital Health 2.10 (2020): e516-e525.

[5] Fine, Paul, Ken Eames, and David L. Heymann. "“Herd immunity”: a rough guide." Clinical infectious diseases 52.7 (2011): 911-916.

[6]Allam, Mayar, et al. "COVID-19 diagnostics, tools, and prevention." Diagnostics 10.6 (2020): 409.

[7] Yan, Li, et al. "An interpretable mortality prediction model for COVID-19 patients." Nature machine intelligence 2.5 (2020): 283-288.

[8] Khadem, Heydar, et al. "COVID-19 mortality risk assessments for individuals with and without diabetes mellitus: Machine learning models integrated with interpretation framework." Computers in Biology and Medicine 144 (2022): 105361.

[9] P. Jonsson, C. Wohlin, An evaluation of k-nearest neighbour imputation using likert data, in: 10th International Symposium on Software Metrics, 2004. Proceedings, 2004, pp. 108–118, https://doi.org/10.1109/ METRIC.2004.1357895.

[10] N. V Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https:// doi.org/10.1613/jair.953.

[11] A.K. Das, S. Mishra, S.S. Gopalan, Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool, PeerJ 8 (e10083) (2020) 1–12, https://doi.org/10.7717/peerj.10083.

[12] A.M.U.D. Khanday, S.T. Rabani, Q.R. Khan, N. Rouf, M.M.U. Din, Machine learning based approaches for detecting COVID-19 using clinical text data, Int. J. Inf. Technol. 12 (3) (2020) 731–739, https://doi.org/10.1007/s41870-020-00495-9.

[13] Rath, Smita, AlakanandaTripathy, and AlokRanjanTripathy. "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model." Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14.5 (2020): 1467-1474.

[14] Yang, Qiao, et al. "Clinical characteristics and a decision tree model to predict death outcome in severe COVID-19 patients." BMC infectious diseases 21.1 (2021): 1-9.

[15] Das, Ashis Kumar, Shiba Mishra, and SajiSaraswathyGopalan. "Predicting CoVID-19 community mortality riskusing machinelearning and development of an online prognostic tool." PeerJ 8 (2020): e10083.

[16] Muhammad, L. J., et al. "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset." SN computer science 2.1 (2021): 1-13.

AUTHOR PROFILE

Rana
received his B.Tech. degree in computer science & engineering from DAV University, Jalandhar, Punjab, India in 2020 and pursuing M.Tech.in computer science & engineering from Sant Baba Bhag Singh University, Jalandhar, Punjab, India.

*e*ISSN1303-5150