P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia
Science & Technology Asia
Vol. 29 No.4 October - December 2024
Page: [51-61]
Original research article

Application of Data Mining in the Prediction of COVID-19 Outcome

Alka Dilip Gore[1,*], Vinayak Jadhav[2], Aniket Muley[2], Sheetu Jailkhani[3], Mayuri Rotti[1], Vivek Waghachavare[1], Randhir Dhobale[1], Girish Dhumale[1]

[1]*Department of Community Medicine, Bharati Vidyapeeth (Deemed to be University) Medical College and Hospital, Maharashtra 416410, India*
[2]*School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Maharashtra 431606, India*
[3]*Department of Community Medicine, Rajarshi Chhatrapati Shahu Maharaj Government Medical College, Maharashtra 416003, India*

Received 2 May 2024; Received in revised form 9 November 2024
Accepted 4 December 2024; Available online 27 December 2024

ABSTRACT

In December 2019, the novel coronavirus, COVID-19, emerged in Wuhan, China, and rapidly spread across the globe, leading to a significant increase in morbidity and mortality rates. The virus presented with diverse clinical manifestations, and robust predictive models were needed to anticipate outcomes and implement timely preventive and corrective measures. This study was designed to identify patterns in COVID-19 outcomes and develop prediction models for patient survival using data mining techniques. The study was conducted at a tertiary care hospital, Bharati Vidyapeeth (Deemed to be University) Medical College and Hospital, Sangli, analysing cases from June 2020 to December 2020. Data were retrospectively collected from the Record Department using a structured pro forma form and analysed using Microsoft Office 2016, SPSS-22, and WEKA-3.8.6, with cases completing at least 80% of the information. Various simple and ensemble machine learning algorithms were applied to classify patient survival and COVID-19 test results. Through statistical and data mining approaches, the study identified patterns in parameters for both survivors and non-survivors, as well as COVID-positive and negative patients. The finalised model for predicting patient survival or non-survival was *functions.SMO*, with 71.64% (±0.83%) of instances correctly classified; and for distinguishing COVID-19 positive from negative cases, the best-performing model was *trees.RandomForest*, achieving an accuracy of 84.41% (±0.35%). These prediction models serve as valuable tools for physicians to diagnose and manage COVID-19, identify critical cases in the early stages, and enhance patient care through timely interventions.

Keywords: Algorithms; Coronavirus; COVID-19; Data Mining; India; Tertiary care hospital

***Corresponding author:** alkadilipgore@gmail.com

1. Introduction

In December 2019, researchers identified a distinct Sars-CoV-2 coronavirus in Wuhan, China. This scourge quickly spread from China to over 100 other nations [1, 2]. The first case in India was reported in Kerala on Jan 27, 2020 [3]. On a global basis, today (March 2, 2021), the number of infected cases is 115,034,561, and more than 2,551,380 death cases were reported, whereas, in India, these figures are 11,124,527 with 157,275 deaths [4].

Corona gradually spread over the globe. The number of cases was not decreasing, and the coronavirus was not abating. It had to be halted somewhere; otherwise, the entire human species would have been demolished. Doctors, nurses, healthcare professionals, and researchers were all scrambling to find a way out of this dilemma.

The World Health Organization (WHO) has identified a range of symptoms associated with COVID-19, which vary in severity and frequency; common symptoms include fever, dry cough, and fatigue, while less common symptoms include headache, sore throat, diarrhea, conjunctivitis, loss of smell, and skin rashes. In more severe cases, COVID-19 can cause breathing problems, chest discomfort, and loss of speech or movement [5]. Several academics developed prediction models for the massive health issues of COVID-19. These are useful for health systems in making strategic decisions, preventing virus propagation, creating social isolation plans to limit the possibility of contagions, and developing pandemic mitigation measures [6]. AI approaches have been extensively applied and evaluated in the healthcare industry, and the recently found COVID-19 requires applying these techniques in recognising, forecasting, and preventing its outbreak [7, 8]. Predictive analytics becomes critical when dealing with a large amount of susceptible data [9]. Data mining techniques have a massive impact on pandemic studies by assisting researchers in revealing the unknown characteristics of a new epidemic and the next future pandemic. Data mining methods have successfully infiltrated our daily lives and contributed to humans' victory in the immensely difficult war against COVID-19 [9].

The application of intelligent and clever innovations, like artificial intelligence, machine learning, and data mining, can be used as an assistance within the early distinguishing proof of potential cases of COVID-19 [10]. Data mining is the complex procedure of extracting valuable information from databases, involving discovering implicit, previously undisclosed, and potentially beneficial knowledge. It is a nontrivial process that aims to uncover hidden insights and patterns within the data stored in a database [11, 12]. Data mining strategies can be applied blindly; statistically, ‘data fishing or p-hacking’ can be unsafe and dangerous, which may lead to the discovery of meaningless patterns. Hence, finding the correct and meaningful data pattern is significant. Laboratory results are accurate, but we may not get them in time due to such laboratories' high costs and low availability. The accuracy of COVID-19 diagnosis must be upgraded to immediately and correctly identify positive patients and provide early therapy [13].

There is a need for rapid and accurate diagnosis based on clinical symptoms and findings. Still, due to a lack of specialised medical equipment and to time-consuming and expensive treatments, it is not always feasible for the ordinary person to receive treatment in time. This research aimed to examine data mining algorithms on the ‘Weka’ tool to design the models and recognise patterns from the data set, so clinicians may begin efforts to cure the sickness as soon as possible.

Hence, the study was performed to develop the prediction model of COVID-19 patients, using data mining with supervised

machine learning algorithms and to compare the pattern of different laboratory parameters and outcomes, such as deaths/discharges and positive/negative COVID-19 patients. The novelty of this study lies in its comprehensive approach to building predictive models using machine learning algorithms that combine both clinical and socio-demographic data. Unlike previous studies, we applied advanced data mining techniques that allowed us to develop highly accurate models for predicting survival and distinguishing COVID-19 positive from negative cases.

2. Materials and Methods

A hospital record-based, retrospective observational and single-centred study of COVID-19 cases was conducted in a tertiary care COVID hospital at Bharati Vidyapeeth Medical College and Hospital, Sangli. Data related to COVID-19 patients, stored at the Documentation group of B.V.D.U.M.C.H, Sangli, was considered for the analysis. The clinical laboratory and radiology departments confirmed the reports as needed. All records of COVID-19-positive, as well as negative patients, were taken for the research.

The Research Review and Institutional Ethical Committee were permitted to conduct the research. (BV(DU)MC&H/Sangli/IEC/439/21)

Data (from June 2020 to December 2020) was collected using the pro forma form prepared by investigators with the help of experts in the field.

Complete records were obtained from the Record Dept. of B.V.D.U.M.C.H, Sangli, and if more than 80% of the information was not written in the record, the case was removed from the analysis. For other cases with minor missing information, imputation methods were applied using the mean for continuous normal variables, median for skewed or ordinal variables, and mode for categorical variables. For some patients, multiple data were recorded for a few factors, but only the last available recording of patients was considered for the analysis.

Following the guidelines provided by Riley RD et al. in their paper on calculating sample sizes for developing clinical prediction models (2020), a sample size of 1338 used in this study is notably not categorised as a small sample size, reflecting a suitable basis for data mining endeavours [14]. Accordingly, data was cleaned and later analysed using appropriate statistical methods, and appropriate software. Data preprocessing was done by normalizing continuous variables, such as laboratory values, to ensure comparability. Outliers were identified and removed based on clinical relevance. Categorical variables, such as socio-demographic factors, were encoded before being fed into machine learning models.

Socio-demographic factors, such as age, gender, and residence from the records, were considered in the study. Symptoms such as fever, headache, body ache, vomiting, cough, common cold, running nose, breathlessness, decreased appetite, sore throat, loose motion, etc., were documented. Records of lab parameters, like Hemoglobin, TLC, Polymorph, Lymphocytes, Platelets, ESR, HIV / HbsAg / HCV, SGOT, SGPT, Bilirubin, Protein, ALP, electrolytes, blood urea, Sr. Creatinine, Sr. Ferritin, BSL, LDH, CRP, D-dimer, IL6, etc., were taken. The study considered 53 different factors with two different outcomes: death/discharge and COVID positive/negative. Data obtained from the record department was in Microsoft Excel format. It was collected and entered according to the pro forma format and then cleaned. After handling the missing values, it was converted into comma-separated value (CSV) format, which is easy to import and analyse in the data mining software- 'WEKA.'

Survival was defined as patients being discharged alive from the hospital,

while non-survival referred to mortality during the hospital stay. Laboratory parameters were assessed based on standard clinical thresholds, and symptoms were recorded according to the hospital's clinical guidelines for COVID-19.

Frequency, percentages of qualitative data, and mean and standard deviation of quantitative data were obtained. A chi-square test was applied to the data to find the association of different factors with outcomes. An unpaired t-test was applied to compare lab parameters between surviving and non-surviving patients and positive and negative patients. Several models were built using the WEKA Explorer module with supervised machine-learning techniques.

Prediction models regarding death/discharge and COVID-19 positive/negative were obtained through the data-mining procedures. Simple and ensemble models were selected for their strong classification performance in previous studies involving similar clinical datasets. Functions.SMO is known for its ability to handle non-linear relationships, while RandomForest offers robustness in handling high-dimensional data and prevents overfitting. Initially, some important simple algorithms, like rules.ZeroR, SimpleLogistic, lazy.IBk, trees.REPTree, trees.J48, functions.SMO, functions.MultilayerPerceptron, bayes.NaiveBayes; and ensemble algorithms like Meta.Bagging trees.RandomForest, meta.AddaBoost, meta.Vote and meta.Stacking were applied to both datasets. A 10-fold cross-validation technique was applied to evaluate the performance of the models and avoid overfitting. Hyperparameters for the trees.RandomForest and functions.SMO models were tuned using grid search to achieve optimal performance.

A few models were chosen based on their classification accuracy, and the final prediction model with high classification accuracy and low standard deviation was chosen by using the experimenter. The effectiveness of complete classification models was assessed using a tenfold cross-validation process method. For data analysis, Microsoft Excel Office 365, SPSS-22, and WEKA 3.8.6 as data mining tools were used.

3. Results and Discussion

3.1 Results

The data from 1338 cases, comprising 406 females (30.34%) and 932 males (69.66%), were studied. The study considered various outcomes, including death or discharge and COVID-positive or COVID-negative results. Among the patient population, 943 individuals (71.1%) were discharged, while 383 (28.9%) unfortunately passed away. Out of the total number of patients, 1128 individuals (84.3%) tested positive for COVID-19, while 210 patients (15.7%) tested negative for the virus.

In July, out of 148 patients, 143 (96.6%) tested positive for COVID-19, and 124 patients (83.8%) passed away, which is significantly higher over six months. ($p < 0.05$)

Many of the patients, 529 (39.54%), were over 60 years old, followed by patients who were 46 to 60 years old (419-31.32%), and the fewest patients (14-1.05%) were under the age of 18. Most patients, 868 (64.87%), came from rural areas. The patients presented a range of symptoms related to COVID-19. Among the cases, 873 patients (65.25%) experienced breathlessness, 689 patients (51.49%) had a cough, 673 individuals (50.30%) had a fever, and 360 patients (26.91%) reported symptoms such as headache or body pain. Besides the symptoms mentioned earlier, patients also presented with other typical manifestations associated with COVID-19, including the common cold, vomiting, sore throat, decreased appetite, loose stools, and runny nose. Major risk factors among these patients were diabetes 421 (31.46%), hypertension 508 (37.97%), chronic renal

disease 61 (4.56%), ischemic heart disease 59(4.41%), and asthma 15 (1.12%). (Table 1)

Table 1. Frequency distribution table of patients by demographic and clinical characteristics.

Patients' characteristics	No. of patients	Percentage
Age Groups (in years)		
<= 5	4	0.3
6 – 18	10	0.75
19 – 25	27	2.02
26 – 45	349	26.08
46 – 60	419	31.32
>= 61	529	39.54
Sex		
Female	406	30.34
Male	932	69.66
Residence		
Rural	868	64.87
Urban	470	35.13
Symptoms		
Head-ache / Body-ache	360	26.91
Fever	673	50.3
Vomiting	70	5.23
Cough	689	51.49
Common Cold	110	8.22
Running Nose	11	0.82
Breathlessness	873	65.25
Decreased Appetite	50	3.74
Sore Throat	57	4.26
Loose Motion	44	3.29
Risk factors		
Diabetes	421	31.46
Hypertension	508	37.97
Chronic Kidney Disease	61	4.56
Ischemic Heart Disease	59	4.41
Asthma	15	1.12

Note: Patients have multiple symptoms and risk factors. therefore, the total count exceeds the total number of patients.

The study observed a significant association between the outcome of death or discharge and various circumstances, such as interaction with COVID-19-positive individuals and using ventilators or oxygen ($p < 0.01$). Additionally, it was found that symptoms like breathlessness and sore throat were also significantly associated with the outcome of death or discharge ($p < 0.05$). Out of the total 13 patients who required a ventilator, 10 individuals (76.1%) did not survive. Among those who did not survive, 274 individuals (31.60%) reported experiencing breathing problems.

The study found a significant association between the outcome (positive or negative) and several symptoms like fever, shortness of breath, and decreased appetite. These symptoms were observed to impact the outcome of the patients substantially. ($p < 0.05$). A significantly large number of patients, 584 (86.8%) out of 673 having fever, were positive. Similarly, fewer patients did not experience breathlessness, with 57 (12.3%) and 197 (15.3%) having decreased appetite, respectively.

Comparisons were made between the means of the outcomes (survival / non-survival and COVID positive/negative) for several continuous variables, including age, the number of days that the patient had symptoms, Hb levels, TLC, polymorphs, lymphocytes, platelets, ESR, Liver Function Tests (LFTs), Sr. Na and Sr. K. The study revealed significant differences in various factors between patients who survived and those who did not. These differences include age, haemoglobin (Hb) levels, polymorphs, serum protein, albumin, C-reactive protein (CRP) levels, and duration of hospital stays. These factors were crucial in distinguishing the outcomes between the two groups ($p < 0.05$). There were statistically significant variations in polymorphs, indirect bilirubin, albumin, and blood urea between COVID-positive and negative patients ($p < 0.05$). Non-survivors and COVID-positive patients were older, had higher SGPT, SGOT, Sr. Creatinine, Blood urea, Blood sugar levels, Ferritin, LDH, CRP D-dimer, and IL6 levels, and lower Hb, platelets, sr. Protein and albumin levels. The mean hospital stay of positive patients was 7.65 ± 5.82 days with a range of (0, 62); for negative patients, it was comparatively less, i.e., 5.93 ± 4.69 with a range of (0, 35) days. Whereas hospital stay for surviving patients was 7.61 ± 5.83 (range: 0, 62), for non-surviving, it was 6.80 ± 5.3 (range 0, 43)

Data were further analysed using data mining tools to find the prediction model for the outcome of death/discharge.

The cross-validation method was applied as a test option since it is more accurate when making predictions (Brownlee, 2017). Since the class of dataset is categorical, the percentage of correctly classified instances, mean absolute error, root mean squared error, ROC area, and confusion matrix were considered to determine the best prediction models. Initially, a few simple algorithms like ZeroR, SimpleLogistic, lazy.IBk, trees.REPTree, trees.J48, functions.SMO, functions.MultilayerPerceptron,

Bayes.NaiveBayes were applied. The models with the highest classification accuracy from simple algorithms were ZeroR with 71.12, simple logistic with 70.89, and support vector regression with 71.64 percent accuracy. (Table 2)

Table 2. Summary of Simple Algorithms for Outcome: Discharge / Death.

Correctly classified instances	Mean absolute error	Root mean squared error	ROC area Discharge	ROC area Death	Confusion Matrix
1. ZeroR					
71.1161	0.411	0.4532	0.496	0.496	943 0 | a = Discharge 383 0 | b = Death
2. Logistic Regression (functions.SimpleLosgistic).					
70.8899	0.429	0.4665	0.55	0.55	920 23 | a = Discharge 363 20 | b = Death
3. k-Nearest Neighbors (lazy.IBk).					
61.3122	0.3871	0.6215	0.509	0.509	706 237 | a = Discharge 276 107 | b = Death
4A. Classification and Regression Trees (trees.REPTree).					
68.6275	0.4059	0.4673	0.546	0.546	881 62 | a = Discharge 354 29 | b = Death
4B. Classification and Regression Trees (trees.J48).					
68.8537	0.3998	0.4761	0.532	0.532	871 72 | a = Discharge 341 42 | b = Death
5. Support Vector Regression (functions.SMO).					
71.644	0.2836	0.5325	0.511	0.511	940 3 | a = Discharge 373 10 | b = Death
6. Artificial Neural Network (functions.MultilayerPerceptron).					
58.5219	0.4084	0.604	0.527	0.527	678 265 | a = Discharge 285 98 | b = Death
7. Naïve Bayes (Bayes. NaïveBayes)					
55.6561	0.4631	0.5734	0.529	0.529	558 385 | a = Discharge 203 180 | b = Death

Ensemble algorithms like meta.Bagging, trees.RandomForest, meta.AddaBoost, meta.Vote and meta.Stacking were employed. The models with the highest classification accuracy used ensemble algorithms Bagging with a classification accuracy of 70.21 (classifier J48, and a numiterations 100), AddaBoost 71.19 (classifier- DecisionStump), voting 71.12 (classifier – ZeroR) and stacking 71.12(for all meta classifiers: ZeroR, J48, and SMO) were finalized. (Table 3)

Table 3. Summary of Ensemble Algorithms for Outcome- Discharge / Death.

	Correctly classified instances	Mean absolute error	Root mean squared error	ROC area Discharge	ROC area Death	Confusion Matrix
1. Bagging (Bootstrap Aggregation) Meta-Bagging (Bag size %= 100)						
Classifier-REPTree	69.6078	0.3936	0.4566	0.583	0.583	887 56 | a = Discharge 347 36 | b = Death
J48, numIterations=100	70.2112	0.3935	0.4521	0.581	0.581	896 47 | a = Discharge 348 35 | b = Death
2. Random Forest (Extension of Bagging) tree-RandomForest						
Numfeatures=0	70.0603	0.4003	0.4488	0.6	0.6	913 30 | a = Discharge 367 16 | b = Death
Numfeatures=10	69.7587	0.3971	0.4512	0.592	0.592	904 39 | a = Discharge

						362 21 | b = Death
3. AddaBoost						
Classifier-DecisionStump	71.1916	0.3954	0.4496	0.593	0.593	924 19 | a = Discharge 363 20 | b = Death
J48	64.1026	0.3666	0.5464	0.564	0.564	747 196 | a = Discharge 280 103 | b = Death
Random forest	68.4012	0.3158	0.555	0.54	0.54	888 55 | a = Discharge 364 19 | b = Death
4. Voting- Meta-Vote						
Classifier-ZeroR	71.1161	0.411	0.4532	0.496	0.496	943 0 | a = Discharge 383 0 | b = Death
SMO	55.6561	0.2836	0.5325	0.511	0.511	940 3 | a = Discharge 373 10 | b = Death
5. Stacked Generalization -meta.stacking						
Metaclassifier-ZeroR	71.1161	0.411	0.4532	0.496	0.496	943 0 | a = Discharge 383 0 | b = Death
J48	71.1161	0.411	0.4532	0.496	0.496	943 0 | a = Discharge 383 0 | b = Death
SMO	71.1161	0.2888	0.5374	0.5	0.5	943 0 | a = Discharge 383 0 | b = Death

The finalized model was the functions.SMO, by using the experimenter, with a classification accuracy of 71.64 (±0.83%).

For another outcome- COVID-19 positive/negative and for categorical class- Cross-validation method with few simple algorithms like ZeroR, SimpleLogistic, lazy.IBk, trees.REPTree, trees.J48, functions.SMO, functions.MultilayerPerceptron, Bayes.NaiveBayes; and ensemble algorithms like Meta.Bagging, trees.RandomForest, meta.AddaBoost, meta.Vote and meta.Stacking was used. The percentage of correctly classified instances, mean absolute error, root mean squared error, ROC area, and confusion matrix were considered to establish the best prediction models. Simple logistic with 83.48, REPTree with 83.26, trees.J48 with 84.30, and support vector regression with 84.3 were the models with the highest classification accuracy from simple algorithms. (Table 4)

Table 4. Summary of Simple Algorithms for Outcome: positive /negative.

Correctly classified instances	Mean absolute error	Root mean squared error	ROC area Positive	ROC area Negative	Confusion Matrix ===
1. ZeroR					
80.3049	0.265	0.3638	0.499	0.499	1128 0 | a = Positive 210 0 | b = Negative
2. Logistic Regression (functions.SimpleLosgistic).					
83.4828	0.2489	0.3633	0.65	0.65	1106 22 | a = Positive 199 11 | b = Negative
3. k-Nearest Neighbors (lazy.IBk).					
70.9268	0.2911	0.5387	0.539	0.539	887 241 | a = Positive 148 62 | b = Negative
4A. Classification and Regression Trees (trees.REPTree).					
83.2586	0.2616	0.37	0.578	0.578	1108 20 | a = Positive 204 6 | b = Negative
4B. Classification and Regression Trees (trees.J48).					
84.3049	0.2636	0.3637	0.502	0.502	1127 1 | a = Positive 209 1 | b = Negative
5. Support Vector Regression (functions.SMO).					
84.3049	0.157	0.3962	0.5	0.5	1128 0 | a = Positive 210 0 | b = Negative
6. Artificial Neural Network (functions.MultilayerPerceptron).					
77.728	0.2279	0.4405	0.649	0.649	997 131 | a = Positive 167 43 | b = Negative
7. Naïve Bays (Bayes.NaïveBays)					
57.6233	0.4284	0.5781	0.538	0.538	676 452 | a = Positive 115 95 | b = Negative

Bagging with classification accuracy 84.23%- classifier REPTree and 84.45%- classifier J48 and numiterations 100, RandomForest 84.45%, AddaBoost 84.3%- classifier - DecisionStump, voting 84.30%- classifier – ZeroR, SMO and stacking 84.3%- for all meta classifiers: ZeroR, J48, and SMO. (Table 5)

Table 5. Summary of Ensemble algorithms for outcome: positive /negative.

	Correctly classified instances	Mean absolute error	Root mean squared error	ROC area Positive	ROC area Negative	Confusion Matrix ===
1. Bagging (Bootstrap Aggregation) Meta-Bagging (Bag size %= 100)						
Classifier-REPTree	84.2302	0.2532	0.3615	0.609	0.609	1124 4 | a = Positive 207 3 | b = Negative
J48, numIterations=10	84.1555	0.2581	0.3671	0.578	0.578	1121 7 | a = Positive 205 5 | b = Negative
J48, numIterations=100	84.4544	0.2542	0.358	0.631	0.631	1128 0 | a = Positive 208 2 | b = Negative
2. Random Forest (Extension of Bagging) tree-RandomForest						
Numfeatures=0	84.4544	0.2474	0.3596	0.608	0.608	1128 0 | a = Positive 208 2 | b = Negative
Numfeatures=10	84.3797	0.2426	0.3595	0.62	0.62	1127 1 | a = Positive 208 2 | b = Negative
3. AddaBoost						
Classifier-DecisionStump	84.3049	0.2579	0.363	0.594	0.594	1128 0 | a = Positive 210 0 | b = Negative
J48	79.2975	0.2091	0.4282	0.585	0.585	1036 92 | a = Positive 185 25 | b = Negative
Random forest	83.7818	0.162	0.4009	0.585	0.601	1114 14 | a = Positive 203 7 | b = Negative
4. Voting- Meta-Vote						
Classifier-ZeroR	84.3049	0.265	0.3638	0.499	0.499	1128 0 | a = Positive 210 0 | b = Negative
SMO	84.3049	0.157	0.3962	0.5	0.5	1128 0 | a = Positive 210 0 | b = Negative
5. Stacked Generalization -meta.stacking						
Metaclassifier-ZeroR	84.3049	0.265	0.3638	0.499	0.499	1128 0 | a = Positive 210 0 | b = Negative
J48	84.3049	0.2646	0.3638	0.499	0.499	1128 0 | a = Positive 210 0 | b = Negative
SMO	84.3049	0.157	0.3962	0.5	0.5	1128 0 | a = Positive 210 0 | b = Negative

By using the experimenter, it was found that meta.Bagging and trees.RandomForest had the same classification accuracy, i.e., 84.41, but different s.d. Hence, the finalized model was trees.RandomForest with minimum variation (± 0.35%).

3.2 Discussion

COVID-19 data was supplied by governments from many countries, corporate sectors, and non-governmental organisations to track epidemics. However, no application has been developed that collects worldwide fine-grained data, such as demographics, environmental conditions, or other covariates, and harmonises the vast quantity of heterogeneous data that has become accessible [15].

If doctors could forecast a patient's outcome, such as death/discharge or positive/negative, using a rapid statistical or data analysis prediction model, they might devote more weightage to critical patients to enhance the survival probability of those patients. As it was a record-based study, the data in this study were not classified uniformly and there was an unbalanced data problem. In the research, there were 943 (71.1%) discharges, 383 (28.9%) deaths, 1128 (84.3%) COVID-positive patients, and

only 210 (15.7%) negative patients. Most of the positive and negative people who had significant symptoms were hospitalised. As a result, the number of negative patients is relatively low.

As a tertiary care center, most patients hospitalised in our COVID unit were moderate to severe. As a result, the case fatality rate was equally high (28.9%). Excess mortality was observed during this pandemic due to various factors, including delayed care, overwhelmed healthcare systems, a shortage of trained healthcare professionals, severe morbidities, etc. Moulaei noticed just 114 (7.6%) deceased patients. He gathered the information from Ayatollah Taleghani Hospital in the southwest Khuzestan Province, Iran [16].

In this study, we have used various simple and ensemble algorithms and found that these models had the accuracy of classifying the instances ranging from 55.65% to 71.64%. The finalised prediction model for deaths/discharges was 'functions.SMO' with a classification accuracy of 71.64 (± 0.83%). Data scientists use the SVM because it may achieve strong generalisation performance without prior information or expertise [5,15,17]. Villavicencio CN et al. (2021) observed that the RF was the high-performing model with a shorter training time, followed by SVM, showing the highest accuracy, 98.81% (± 0.012) [5].

Our analysis showed that RandomForest had the highest classification accuracy of 84.41 (0.35%). In contrast, Ibrahim Arpaci (2021) observed that the CR meta-classifier had the most accurate classifier, with an accuracy of 84.21% [8]. Moulaei observed the RandomForest method as the most excellent performance algorithm, with an accuracy of 95.03% [16]. The prediction accuracy of the analysis of COVID-19 positive and negative instances was roughly the same as Arpaci but lower than Moulaei. Sadig et al. (2020) used the J48 classification algorithm to classify positive and negative patients through the 'WEKA' tool to differentiate COVID cases between positive and negative [18]. They showed prediction accuracy (62%).

Brinati, Davide Campagner, et al. developed two machine learning models for the exact prediction with accuracy ranging from 82% to 86% [19].

Albahari performed a systematic review and thoroughly analysed automated AI applications based on data mining and machine learning techniques for identifying and diagnosing COVID-19 in 2020 [20]. He noted that k-NN, Decision tree, and naïve Bayes models with 90% accuracy were used to improve infection prediction in some studies., and some studies found naïve Bayes and J48 with an accuracy of 53.6% and 71.58% were used to build prediction models. Their investigation discovered that overall accuracy ranged from 53.6 to 90%. The accuracy of predicting COVID-19 positive or negative patients was reported differently in different studies. It might be because of different characteristics, a different number of instances, or a different number of attributes considered for modelling.

Although RT-PCR is the gold standard for confirming COVID-19 infection, it is hampered by a shortage of reagents, is time-consuming, and needs specialised facilities. As a result, a less expensive and quicker diagnostic model is required to distinguish between positive and negative COVID-19 patients and categorise the critical patients.

The findings might aid in the early detection of COVID-19 and other emerging diseases, particularly when specialised equipment is insufficient to detect the infection. The study and the prediction models obtained might help primary care physicians save lives by offering in-depth knowledge of this disease and existing medical diagnosis techniques for this virus.

4. Conclusion

Using statistical and data mining procedures, we identified patterns in various clinical and laboratory parameters associated with survival and non-survival patients and COVID-positive and negative patients. Based on these findings, we developed predictive models, which accurately assess patients' outcomes. These models are valuable for physicians to identify critical patients in the early stages, and facilitate timely interventions, which can improve patient care and resource allocation.

The models developed for predicting patient survival and distinguishing COVID-19 positive from negative cases demonstrated robust performance. Both models showed high accuracy, with functions.SMO demonstrated 71.64% accuracy and trees.RandomForest achieved 84.41%. The functions.SMO model for survival prediction and the trees.RandomForest model for classification of COVID-19 cases offers practical tools for clinicians, enabling early detection of critical patients and informed decision-making.

One important limitation of the study is the retrospective design, which may introduce bias due to incomplete or missing data, although efforts were made to minimise this bias. Additionally, the study was conducted in a single centre, which may limit the generalizability of the findings to other populations.

Acknowledgements

We are grateful to the administration of Bharati Medical College and Hospital, Sangli, for permitting us to use the data on patients with COVID-19. We are thankful to all the volunteers who assisted in data collection and entry as a part of a unified fight against COVID-19.

References

[1] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33.

[2] De Ceukelaire W, Bodini C. We need strong public health care to contain the global corona pandemic. Int J Health Serv. 2020 July 1;50(3):276-7.

[3] Andrews MA, Areekal B, Rajesh KR, Krishnan J, Suryakala R, Krishnan B, et al. First confirmed case of COVID-19 infection in India: A case report. Indian J Med Res. 2020 May 1;151(5):490-2.

[4] Anonim. Worldometer. p. 1-25; 2022. COVID live – coronavirus statistics – Worldometer [cited May 10 2022]. Available from: https://www.world-ometers.info/coronavirus/.

[5] Villavicencio CN, Macrohon JJE, Inbaraj XA, Jeng JH, Hsieh JG. Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using Weka. Algorithms. 2021;14(7).

[6] Cortés-Martínez KV, Estrada-Esquivel H, Martínez-Rebollar A, Hernández-Pérez Y, Ortiz-Hernández J. The state of the art of data mining algorithms for predicting the COVID-19 pandemic. Axioms. 2022;11(5):242.

[7] Bian J, Modave F. The rapid growth of intelligent systems in health and health care. Health Informatics J. 2020;26(1):5-7.

[8] Arpaci I, Huang S, Al-Emran M, Al-Kabi MN, Peng M. Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. Multimed Tools Appl. 2021 March 1;80(8):11943-57

[9] Mengistie TT. COVID-19 outbreak data analysis and prediction modeling using data mining technique. Int J Comput (IJC). 2020;38(1):37-60.

[10] Srinivasa Rao ASR, Vazquez JA. Better hybrid systems for disease detections and early predictions. Clin Infect Dis. 2022 February 1;74(3):556-8.

[11] Çela EK, Frasheri N. A literature review of data mining techniques used in healthcare databases. ICT Innov. 2012 Web Proceedings; 2012:577-82.

[12] Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. p. 82-8; 1996. International Conference on Knowledge Discovery and Data Mining [cited May 10 2022]. Available from: http://www.aaai.org/Papers/-KDD/1996/KDD96-014.

[13] Peng M, Yang J, Shi Q, Ying L, Zhu H, Zhu G, et al. Artificial Intelligence Application in COVID-19; Diagnosis and Prediction. SSRN Journal. 2020 (April).

[14] Riley RD, Ensor J, E Snell KI, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. 2020 [cited 2023 Apr 13]; Available from: http://www.bmj.com/-permissionsSubscribe:http://www.bmj.com/subscribeBMJ2020;368:m441doi:10.1136/bmj.m441

[15] Guidotti E, Ardia D. COVID-19 data hub. J Open Source Softw. 2020 July 10;5(51):2376.

[16] Moulaei K, Shanbehzadeh M, Mohammadi-Taghiabad Z, Kazemi-Arpanahi H. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inform Decis Mak. 2022;22(1):2.

[17] Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. IEEE Trans Neural Netw. 1999;10(5):1055-64.

[18] Al Sadig M, Khalid N, Sattar A. Developing a prediction model using J48 algorithm to predict symptoms of COVID-19 causing death. Int J Comput Sci Netw Secur. 2020;20(8):80-3.

[19] Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. J Med Syst. 2020;44(8):135.

[20] Albahri AS, Hamid RA, Alwan JK, Al-Qays ZT, Zaidan AA, Zaidan BB et al.. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): a Systematic Review. J Med Syst. 2020;44(7):122.