Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Mortality prediction of COVID-19 patients using soft voting classifier

Nishant Rai, Naman Kaushik, Deepika Kumar*, Chandan Raj, Ahad Ali

Department of Computer Science & Engineering, Bharati Vidyapeeth’s College of Engineering, New Delhi 110063, India

Abstract

COVID-19 is a novel coronavirus that spread around the globe with the initial reports coming from Wuhan, China, turned into a pandemic and caused enormous casualties. Various countries have faced multiple COVID spikes which put the medical infrastructure of these countries under immense pressure with third-world countries being hit the hardest. It can be thus concluded that determining the likeliness of death of a patient helps in avoiding fatalities which inspired the authors to research the topic. There are various ways to approach the problem such as past medical records, chest X-rays, CT scans, and blood biomarkers. Since blood biomarkers are most easily available in emergency scenarios, blood biomarkers were used as the features for the model. The data was first imputed and the training data was oversampled to avoid class imbalance in the model training. The model is composed of a voting classifier that takes in outputs from multiple classifiers. The model was then compared to base models such as Random Forest, XGBoost, and Extra Trees Classifier on multiple evaluation criteria. The F1 score was the concerned evaluation criterion as it maximizes the use of the medical infrastructure with the minimum possible casualties by maximizing true positives and minimizing false negatives.

1. Introduction

The COVID outbreak of 2019 was the biggest crisis that humanity has faced in recent history. A virus termed Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was the cause of this outbreak. It is one of the members of the coronavirus family of viruses and was named so in January 2020. As of April 2022, the World Health Organization (WHO) has already reported 6.2 million confirmed deaths caused by COVID infection. This virus was first detected in Wuhan, China on December 31, 2019, as reported by the World Health Organization (Roberts et al., 2021). The cases of the same COVID-19 infection were identified outside China through early 2020. By March 2020, the reports were growing exponentially from around the world, when the World Health Organization classified the outbreak as a pandemic. The quick propagation of this virus all over the world resulted in huge demand for medical resources such as medicines, hospitals, and medical staff. Enormous efforts were made by governments all around the world, such as imposing lockdowns, necessitating isolation of infected patients and creating awareness about precautionary measures like masks and social distancing. Despite these measures, multiple countries have been hit by as many as three or four COVID waves, where a “wave” denotes a spike in COVID infections concentrated over a very small-time frame (Emanuel et al., 2020). The pandemic has hit the medical infrastructure of every country hard and the situation becomes worse in third-world countries where the medical infrastructure is not optimal to handle normal circumstances (Malik, 2022). Since the medical system is already under pressure, every new patient creates immense pressure on the hospital staff to manage the patient with existing resources. This decides the severity of the infection and more probability of mortality. The mortality rate of patients during a particular time period in a hospital can be calculated by comparing the average number of deaths to the total number of patients admitted in the given time frame (Wang et al., 2020). During such a medical emergency, the foremost task is to manage the patients since medical supplies and staff are limited. This can be accomplished by identifying high-risk patients earlier during the infection period and delivering the required therapy (Li et al., 2020). According to several studies, COVID-19-infected individuals originally exhibited only moderate symptoms such as cold or cough, but their medical condition quickly deteriorated, resulting in death in many cases (Sohrabi et al., 2020; Hu et al., 2020). During the initial COVID-19 outbreak around the globe in early 2020, Europe was severely hit and the medical infrastructure was overwhelmed by a constantly increasing number of patients (Iftimie et al., 2021). Most of the patients who succumbed to COVID infection exhibited only mild symptoms, such as fever, and cough at the time they were admitted to the hospital. The development of a model to predict the mortality of a patient early after admission is critical and might save many lives by determining the probability of the patient’s death by assessing his/her current condition.

This topic has been a major point of research for the past couple of years due to the role it plays in helping the medical infrastructure to

* Corresponding author.
E-mail address: deepika.kumar@bharatividyapeeth (D. Kumar).

https://doi.org/10.1016/j.jicce.2022.09.001
Received 9 June 2022; Received in revised form 7 September 2022; Accepted 11 September 2022
Available online 13 September 2022
2666-3074/© 2022 The Authors. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
manage the “waves” of COVID. Various techniques have been used in multiple studies to predict the severity and mortality of patients ranging from their race, chest CT scans, medical biomarkers, and past medical records. Symptoms of COVID-19 can help in determining the severity of infection in patients who tested positive for COVID, but at the same time, the symptoms are similar to other common diseases which makes it difficult to diagnose the illness as COVID. Several studies have used machine learning models based on various factors. Some of them used chest X-rays and CT scans to predict mortality by analyzing the condition of the lungs of the patient since it has been established that COVID-19 affects the lungs and causes casualty (X-Lassau et al., 2021). This provides somewhat accurate results but is not feasible at a large scale in a situation where the influx of patients is overwhelming the medical infrastructure. In this case, using the important biological biomarkers and past medical history of the patient obtained from blood tests is much more feasible as it can be performed on a large scale. Biological biomarkers are the important molecules found in the human body that might indicate the presence of a disease or abnormal behavior (Yan et al., 2020). Thus, it is crucial to devise methods that can use other factors and tests to determine the severity of the infection which we aim to achieve with this research.

The past research has been focused on improving the accuracy and F1 score also plays an important evaluation criterion, since it highlights the number of true positives and true negatives predicted by the model. As there are limited ICU beds available and the medical infrastructure capacity is at its limit during the COVID wave, predictive decisions can be made to maximize the number of patients that can be treated in the limited amount of resources available saved for severe patients and thus this research focuses on the F1 score and it has been taken as the main evaluation criteria. The proposed model predicts if the patient admitted to the hospital survives or dies during the hospitalization period. This research proposes a soft voting classifier which uses an ensemble of different classification algorithms viz. Random Forest, XGBoost, Gradient Boosting and Extra Tree Classifier. The ensemble model was then evaluated against the base models on various evaluation criteria such as accuracy, precision, recall, F1 Score, ROC-AUC and PRC-AUC.

The research contributions have been summarized as follows:

- The model proposed in the study determines the probability of death of a patient admitted to a hospital considering various aspects of the patient such as age and the blood biomarkers.
- The proposed model uses a soft voting classifier to predict the mortality of the patient admitted to the hospital with base estimators such as Random Forest, XGBoost, Gradient Boosting Classifier, and Extra Tree Classifier.
- Accuracy, Precision, Recall, and F1 Score have been used as the evaluation criteria for checking the robustness of the proposed methodology.
- Comparing the proposed model with base models and the existing state-of-the-art models provides a superior F1 score which is crucial during the COVID waves.

The paper follows the following structure. In Section 2, earlier work on mortality due to the COVID-19 virus is analyzed. Section 3 discusses the structure of the dataset used to train and test the models, the data cleaning, and feature selection. It is followed by the description of base models and the architecture of the proposed model and the methodology used to create the model. Section 4 displays the confusion matrices and the scores attained by the model and a comparison with other base models.

2. Literature review

A study was performed to identify risk factors associated with the mortality of patients in Wuhan, China, infected with SARS-CoV2. This study consisted of 191 patients whereas, 137 survived and 54 died in the hospital. Some of the risk factors are identified as older age, higher Sequential Organ Failure Assessment (SOFA) score, and D-dimer levels greater than 1.0 µg/mL at the time of admission. Viral shedding was noted to have a duration of 8 days to 37 days and the median duration was 20 days which persisted until death in fatal cases. It was concluded that the risk factors identified in the study could help clinicians identify infected patients with poor prognosis (Zhou et al., 2020). Blood samples from 485 patients from Wuhan, China, were analyzed and XGBoost model was proposed in Chen and Guestrin (2016). The XGBoost multi-tree model ranked the features by their importance and then, the top three key features (biomarkers) of blood samples were chosen: high-sensitivity C-reactive protein, lactate dehydrogenase (LDH) and lymphocytes. Patients at general, severe and critical stages were examined and the outcome (alive or dead) was recorded. Mortality was then predicted ten days before the outcome of the patient (Yan et al., 2020). The same dataset was analyzed in a different study by comparing various machine learning algorithms (XGBoost, random forests, SVM, neural networks, logistic regression, and decision trees). Blood biomarkers such as percentage of neutrophil lymphocytes, LDH, C-reactive protein (CRP), and age were found to be most important based on XGBoost feature selection. A neural network was trained on these five features and an accuracy of 96% was reported (Karthikeyan, Garg, Vinod, and Priyakumar, 2021). A scoring system was developed for COVID-19, in which patients were retrospectively analyzed and the clinical data were compared. This study aimed at dividing patients into low and high-risk groups. LASSO regression and multivariate analysis were employed for the task of scoring patients into low and high-risk groups. This study, however, only analyzed 113 severe patients and some blood biomarkers like lactate dehydrogenase were missing from the clinical data (Shang et al., 2020). For the prediction of mortality among patients being treated in ICU, patient data from Denmark and UK were combined to train mortality prediction models which resulted an AUC of 0.906 at the time of diagnosis, 0.818 at the time of hospital admission, and 0.721 at the time of admission of patient in the ICU. Two models based on XGBoost and one based on logistic regression were trained and validated where age, Body Mass Index (BMI) and hypertension were identified as the common risk factors associated with higher mortality (Jimenez-Solem et al., 2021). A different study focused on the respiratory failure of the patient has been studied. The study used the XGBoost model for training, the model achieved a high mean accuracy value of 0.919 and AUC of 0.77. In the study age, emergency severity index level, respiratory rate, serum lactate, and demographic characteristics played an important role in predicting the respiratory failure of the patient (Bolurani et al., 2021).

Clinical data from 361 patients in Wuhan, China and 106 patients in three Korean medical institutions were combined for a study. The data contained 30 variables which included 28 blood biomarkers and 2 demographic variables, age and gender. Two models, an artificial neural network (ANN) and a convolutional neural network (CNN) have been proposed for the prediction of mortality at an early stage. On normalized data, the ANN had an accuracy of 80% and the CNN had an accuracy of 73%. Age, gender, and the number of lymphocytes and neutrophils present in the blood were found to be the most important factors to determine the mortality of the patient (Lin et al., 2021). A study has been conducted utilising clinical data of patients on 73 blood biomarkers. XGBoost feature selection has been used to determine the LDH, lymphocyte and hs-CRP. A decision tree model has been trained on these three features and the proposed model achieved an accuracy of 90% (Ko et al., 2020). A study has been conducted on 4098 patients admitted to five hospitals in New York, using their electronic health records. Mortality probability after 3, 5, 7 and 10 days of admission of the patient has been reported to the hospital. The proposed XGBoost model outperformed other baseline models having a ROC-AUC of 0.84 at 10 days for mortality prediction. The major factors reported to be of importance for mortality prediction were age, anion gap and CRP (Vaid et al., 2020).

An approach using XGBoost on MIMIC dataset has been used to predict mortality in COVID, pneumonia, and mechanically ventilated patients. It outperformed previous methods of mortality prediction and...
focused on the significance of using machine learning models in clinical environments. The model achieved more positive values of AUROC as the time window came close from 72-h prior prediction to 12-h prior prediction. However, it was notable that the accuracy of the said model trained on 114 COVID patient data was 80.9% with a recall of 0.826 for the 12-h prior prediction. The limitation of this study was that vital signs and laboratory data were required at regular intervals (Ryan et al., 2020). For early prognosis of mortality, a cohort study on patients in Korea, using their socio-demographic data was done during the first wave of the pandemic but the dataset did not contain any laboratory or radiology data. LASSO and SVM models with radial basis kernel function performed best on the dataset. The study predicted mortality at different stages of the infection, early prediction, after 14 days, and after 30 days. The results show good prediction rates in all the cases. The major features were the age and medical history of the patient. If the patient was on medication for diabetes or cholesterol had a major effect on the prediction result (An et al., 2020). Another cohort study based on data from infected and admitted patients in the Netherlands saw promising dataset improvement with more than 80 features including pre-morbid factors like age, gender, occupation, and medical history. Most important features were filtered out using SHAP (SHapley Additive exPlanations) values. An XGBoost model and an LR model were then used to do binary classification on the dataset where the binary value represented the likelihood of mortality after 21 days of infection. In this study, it was also found that those with age > 80 and 70 < age < 80 were at more risk as compared to others. The patients are more prone to mortality after 21 days if they are already on medication for some other disease (Ottenhoff et al., 2021). In studies (An et al., 2020; Ottenhoff et al., 2021), it can be observed that feature selection is necessary and it can reflect the real-life variables that are important before and during the treatment of the patient. In the studies, age, medication, and data known for other diseases increased the AUC. However, more features may increase accuracy and ROC-AUC but would fail to generalize if the datasets contain imbalances and missing values. A combination of sociodemographic data and laboratory data, along with appropriate feature selection, could provide more robust predictions. In a recent study, authors utilized blood biomarkers to train on machine learning models, among which, the logistic regression model achieved the best accuracy of 91% validated against an external dataset. An ALDCC score, scoring age, lymphocyte count, D-dimer, CRP, and creatinine on a scale of 0-100 has been calculated, which was proposed to predict the risk of mortality in COVID patients. This conclusion was based on the performance of individual features on the Logistic Regression model (Rahman et al., 2021).

Analysis of CT scans and Chest X-ray images is a popular approach for the classification of COVID-19 infection in the lungs. A study was conducted to classify COVID-19 using lung CT scan images. GLCM (Gray-Level Co-occurrence Matrix) has been proposed to calculate features for ELM (Extreme Learning Machine), a single hidden layer neural network algorithm. The evaluation of this approach was done using K-F cross-validation with ten-fold cross-validation. Mean accuracy of 76%, a mean precision of 0.77, and a mean F1 score of 0.75 was reported for ten runs of ten-fold cross-validation. The authors reported that the model had good stability, which is critical to making accurate predictions (Gao, 2021). Application of deep neural networks and CNNs (convolutional neural networks) combined with other approaches have also been researched for the task of COVID-19 classification. In such a study on Chest X-ray images of patients, authors used DCN (Depth-wise Separable Convolution) algorithm to derive DCNN (Depth-wise separated Convolutional Neural Network) framework, which was then applied to three CNNs, LeNet-5, VGGNet-16 and ResNet-18. The authors also use the Dilated Convolution algorithm to build the DDCNN (Dilated Depth-wise separated Convolutional Neural Network) framework and apply it to two CNNs, VGG-16 and ResNet-18. Overall accuracy reported for the DCNN and DDCNN framework were 93.26% and 92.85% respectively (Li et al., 2021). Another study using Chest X-ray images utilized U-Net with ResUNet backbone has been conducted to segment the images, then features are extracted using CNN. These features were then used for classification in an ensemble model, which is evaluated using five-fold cross-validation. An accuracy of 99.35 and an F1 score of 0.99 was reported by the authors (Kalaiyanvi & Seetharaman, 2022). Chest X-Ray images combined with clinical data of the patient were also utilized to predict mortality, where chest X-Ray images were processed using Resnet-50 along with a vision transformer. The clinical data washed uses by a feed-forward neural network. Results of image analysis and clinical data were then combined to get the final prediction. This method had an accuracy of 73.2% for predicting mortality in patients admitted to ICU. The major parameters in the prediction were found to be age, cardiovascular disease, creatinine, and white blood cell count (Cheng et al., 2022). A dataset containing clinical data of 4,711 patients was made publicly available and data analysis was also conducted, wherein univariate and bootstrap analysis revealed the independent and dependent features in the dataset. A severity score was also proposed in the same paper. Previous studies have indicated that blood biomarkers can be utilized for mortality prediction in COVID patients. The common challenges faced by authors were imbalance in the dataset and missing values in clinical data. In the next section, we propose a methodology for mortality prediction in COVID-19 patients on a dataset provided in the paper (Altschul, 2021).

3. Research methodology

This research focuses on improving the accuracy and F-measure of mortality prediction of patients infected with COVID-19. This study proposed an ensemble soft voting classifier which utilizes an ensemble of machine learning algorithms. Before providing input to the model, data preprocessing was performed, followed by feature selection on raw data as obtained from the paper (Altschul, 2021). Fig. 1 depicts the flow diagram of the proposed ensemble model employing a soft voting classifier.

3.1. Dataset

The dataset used for the research has been taken from a publicly available study conducted on 4711 patients admitted to the hospital from March 1st, 2020 to April 16th, 2020. The dataset contains various attributes about patients such as patient ethnicity, age, blood biomarkers etc. (Buck, 1960). The “Death” attribute in the dataset represents the condition of all hospitalized patients. “0” and “1” represent if the patient was alive or dead after hospitalization respectively. It has 1148 instances of “death” outcomes and 3563 are of “alive” outcomes. There are various features included in the dataset, several represent ethnicity, a feature represents the age of the patient and many features describe the blood biomarkers obtained from clinical tests. The features that represent the medical condition include the records from the patient’s past medical record and the tests conducted on the patient while they were hospitalized. This data was collected from two different hospitals and is considered a feature represented by the attribute “Derivation Cohort”.

3.2. Data preprocessing

This section details the methods used for cleaning the data and employed data transformation methods. Medical data of the patients are generally incomplete and have many missing values. Moreover, the dataset was imbalanced with a death rate of 24.36%. The dataset had three columns for each feature: a value, threshold, and availability of the feature. The other two columns for each feature except the feature value were dropped as they did not contribute to the mortality prediction. The missing values were required to be imputed since missing values reduce the performance of the model. The dataset contains string values for the column “age”. Machine learning models work best with integer and float values. The other columns were checked for “inf” and

N. Rai, N. Kaushik, D. Kumar et al.
International Journal of Cognitive Computing in Engineering 3 (2022) 172–179

174
“NaN” values. All zero values in columns with real positive values representing blood biomarker records were converted to NaN for multivariate imputation (Buck, 1960). Iterative imputer from the sklearn library (Pedregosa et al., 2011) was then used to handle the “NaN” values in the dataset, to make the dataset more consistent. Then, the dataset was split into 70% training data and 30% test data for experimental analysis. After splitting the dataset into train and test sets, the train set was oversampled using SMOTE (Chawla et al., 2002) oversampling method.

3.3. Feature selection

An embedded method of feature selection has been opted for conducting this research. The motive for using embedded methods is due to their highly accurate, generalizable and interpretable predictions. The random forest also works on the same principle. It evaluates features which have the most influence on the outcome. Therefore, an untuned Random Forest Classifier (RFC) was trained and then used to get the relative importance of the features. Random forests are made up of 400–1200 decision trees, each of which is composed of a random mix of the dataset's data points and features. It is relatively immune to overfitting since not every tree gets all the data or all of the features to work. Every tree has a series of binary questions based on various numbers of attributes. The tree partitions the dataset into two buckets at each node (this is at each question), each of which contains observations that are more similar to one another than to those in the other bucket. As a result, the value of each feature is determined by how “pure” each bucket is. So, the RFC model was trained on the data and the importance of each feature was plotted. The top 22 features having high importance were selected which were: Derivation cohort, LOS, Age,1, OsSats, Temp, MAP, Ddimer, Plts, INR, BUN, Creatinine, Sodium, Glucose, AST, ALT, WBC, Lympho, IL6, Ferritin, CrcProtein, Procalcitonin, Troponin. These features have been named as they appear in the dataset (Altschul, 2021).

3.4. Proposed methodology

In this study, a soft voting ensemble classifier has been proposed for the prediction process. It is a machine learning strategy that combines the outputs produced by various models called base estimators. Since machine learning algorithms have limits and constructing a high-accuracy model is challenging and a single method may not be able to deliver the most accurate prediction for a given dataset. So, multiple models can be used to improve overall accuracy. The combination was created with two goals in mind: lowering model error while retaining generalization. The proposed model was built using the voting classifier. A voting classifier is a machine learning model which uses various other models for classification and uses the method of voting to produce an output (i.e class) based on the class with the highest likelihood of becoming the output.

3.4.1. Model architecture

1. Random Forest: Random forest is a well-known supervised machine learning technique. It is based on the idea of ensemble learning. It aggregates different numbers of decision trees on different subsets of the provided dataset and then averages the results to improve the dataset's predicted accuracy.
2. XGBoost: Boosting is an ensemble learning strategy for creating a strong classifier from a sequence of weak ones. Dealing with the bias-variance trade-off necessitates the use of boosting techniques. XGBoost was trained with gamma set as 0.2, learning rate as 0.1 and maximum depth set as 12.
3. Gradient Boosting: It is a type of ensemble learning. It uses a gradient boosting approach to learning from the induced regression trees. It was trained with a learning rate of 0.2.
4. Extra Tree Classifier: Extra Tree Classifier is an ensemble learning technique that predicts an outcome by merging the outcomes of several de-correlated decision trees gathered in a “forest”.
5. Soft Voting Classifier Ensemble: Voting Classifier supports two types of voting, hard voting and soft voting (Ruta & Gabrys, 2005). The projected output class in hard voting is the one with the most votes, i.e. the one having the highest likelihood of being predicted by each of the classifiers. The hard voting may be equally weighted or unequally weighted which allows assigning higher priority to one of the classifiers. The output class is predicted in soft voting based on the probability assigned to the classes by the classifier. In equally weighted soft voting, all the classifiers have equal importance whereas, in unequally weighted soft voting, different classifiers are assigned different weights depending upon their importance in the classification. The probabilities returned by the classifiers are multiplied by the weights to assign a probability to the classes and the class with the highest value is chosen. The voting type used were soft estimators viz. Random Forest, XGBoost, Gradient Boosting Classifier, and Extra Tree Classifier as these models achieve highest accuracy when trained separately. Random forest was used as it balances error in class-unbalanced datasets with low bias and works well with high-dimensional data. Boosting models were used as they offer high predictive accuracy, a wide range of flexibility, can optimum on many loss functions, and offer multiple options for hyperparameter. The Extra Trees classifier's methods assist in lowering both bias and variance.
4. Results and analysis

This section describes the experimental outcomes of training the models on the dataset, as well as a comparison of the proposed model to other classification methods. The experimental setup included Google Colab environment with Intel Xeon CPU and 13GB RAM. The libraries used were pandas, NumPy, sklearn, imbalanced-learn and xgboost. Data cleaning and preprocessing methods were first applied to the dataset, after which a total of 22 blood biomarkers were recognized as the most important for predicting mortality. Models were trained and tested on 70% and 30% of the data respectively. Baseline models, K-Nearest Neighbor Classifier, Naive Bayes Classifier, Random Forest Classifier, XGBoost Classifier, Gradient Boosting Classifier, and Extra Tree Classifier were trained, and the test results were recorded.

4.1. Evaluating baseline models and proposed methodology

Voting Classifier with ‘soft’ voting was used to build the ensemble model. Baseline classification models with better results were chosen as estimators for the Voting Classifier i.e. Random Forest, XGBoost, Gradient Boosting Classifier, and Extra Tree Classifier. To achieve optimal hyperparameters of the selected estimators, hyperparameter tuning was done using the GridSearchCV method from the Sklearn library (Pedregosa et al., 2011). Accuracy was used as a performance measure in the grid search. A combination of hyperparameters was used as a parameter grid for each estimator. The best combination of parameters obtained after grid search is specified in Table 2.

The models were then compared based on Accuracy (1), Precision (2), Recall (3), and F1 score (4).

\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (1)
\]

\[
\text{Precision} = \frac{tp}{tp + fp} \quad (2)
\]

\[
\text{Recall} = \frac{tp}{tp + fn} \quad (3)
\]

\[
F1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)
\]

Receiver Operator Curves (ROC) for baseline models and the proposed model were compared and their Area Under the Curve (AUC) was represented. Since the dataset contains imbalance, Precision Recall Curves (PRC) of the models were compared and PRC-AUC was also computed. Table 3 depicts the comparative analysis of various models:

Table 3 compares the results of baseline models and the proposed ensemble model for various evaluation metrics. The results show that the proposed soft voting ensemble voting classifier achieved an accuracy of ~87%. Fig. 2 visualizes the results on a line graph. An improvement of 1.34% in accuracy was observed in the proposed ensemble model when compared to other baseline models. Further improvements were observed in precision, F1-score, ROC-AUC and PR-AUC. Fig. 3 displays the confusion matrix of the ensemble model on the test data. The F1 score is a combination of precision and recall which depict the number of true positives compared to the predictions made by the classifier and the number of true positives compared to the true values in the dataset respectively. F1 score is the main focus of the study since it helps maximize the true positives and minimize the false negatives. The confusion matrix in Fig. 3 represents the output of the ensemble classifier where 1 represents death as an outcome whereas 0 represents the patient surviving the COVID infection.

Fig 4(a-d) illustrates the Receiver Operator Curve (ROC) of XGBoost, GradientBoosting, Random Forest and the ensemble soft voting classifiers respectively. Fig 5 compares the Precision Recall Curve (PRC) of the proposed ensemble model, XGBoost model, Random Forest model and no-skil (yes/no classifier). Area Under the Curve of ROC (ROC-AUC) and Area Under the Curve of PRC (PRC-AUC) for the ensemble model found to be the highest among the trained models. It can be observed that even with the imbalance in the test data, AUC-ROC and PRC-AUC were higher for the proposed methodology.

| Table 1
| Description |
|-------------|
| Alive | Dead |
| 3563 | 1148 |

Table 2

| Best Hyperparameters of base estimators after Grid Search. |
|-------------|
| Random forest | XGBoost | Gradient boosting | Extra trees |
| n_estimators = 500 | gamma = 0.2 | n_estimators = 200 | max_features = 10 |
| max_depth = 500 | max_depth = 12 | learning_rate = 0.1 | n_estimators = 500 |
| random_state = 42 | random_state = 42 | random_state = 0 | random_state = 0 |
Table 3
Comparative analysis of various models.

| Models            | Accuracy (%) | Precision | Recall | F1-Score | ROC-AUC | PRC-AUC |
|-------------------|--------------|-----------|--------|----------|---------|---------|
| KNN               | 70.934       | 0.419     | 0.580  | 0.487    | 0.714   | 0.475   |
| Naive Bayes       | 80.198       | 0.594     | 0.527  | 0.558    | 0.816   | 0.565   |
| Extra Trees Classifier | 84.795 | 0.693     | 0.646  | 0.669    | 0.881   | 0.747   |
| XGBoost           | 84.936       | 0.680     | 0.690  | 0.685    | 0.889   | 0.776   |
| Gradient Boosting | 85.219       | 0.682     | 0.708  | 0.695    | 0.886   | 0.769   |
| Random Forest     | 85.643       | 0.698     | 0.696  | 0.697    | 0.882   | 0.735   |
| Ensemble Model    | 86.987       | 0.744     | 0.690  | 0.716    | 0.895   | 0.782   |

Fig. 4. (a-d): ROC curves for various models.

Fig. 5. PRC comparison of proposed ensemble soft voting classifier, XGBoost and Random Forest and no-skill(yes/no classifier). The AUC of the PRC for each model is provided in the legend.

Some key biomarkers including oxygen saturation, Mean Arterial Pressure (MAP), and C-Reactive protein were among the top 22 features chosen. Oxygen saturation is a measurement of how much haemoglobin is currently linked to oxygen vs how much haemoglobin is unbound, and it’s a crucial aspect of patient care management and understanding. MAP is a vital measurement that analyses blood flow across major organs by accounting for flow, resistance, and pressure inside the arteries. C-reactive protein (CRP) is an annular-pentameric protein that is detected in blood plasma and whose levels rise in response to inflammation which can be caused by infection, injury, or chronic disease. Other crucial features included age and body temperature, which can be used to assess a COVID-19 patient.

4.2. Analysis using state-of-the-art models

Table 4 tabulates other techniques used for predicting mortality among COVID-19 patients and their reported results. These studies utilize different datasets and evaluation criteria are being compared separately. A Dataset consisting of 114 patients who tested positive for COVID was trained and evaluated using 5-fold K-fold cross-validation on a hyperparameter-tuned XGBoost model (Ryan et al., 2020). A com-

177
bination of data from 361 patients from Wuhan, China and 106 patients from Korean medical institutions was utilized to train and test an artificial neural network in paper (Lin et al., 2021). With data obtained from five hospitals in New York City for 4098 Covid-19 patients, Extreme Gradient Boosting (XGBoost) was utilized to predict in-hospital mortality at time intervals of 3.5,7, and 10 days from admission (Vaid et al., 2020).

On a dataset containing 2273 patients from the Netherlands, linear logistic regression and non-linear gradient boosting were used to outperform other age-based models at the time (Ottenhoff et al., 2021). A combination of longitudinal chest X-Rays and clinical data of 654 patients was utilized. X-Ray images were analyzed using R50-VIT (Resnet-50 + Vision Transformer) and clinical data was analyzed using a fully connected deep neural network with softmax activation to generate probability (Cheng et al., 2022). 78 clinical features were recorded from 220 patients in Wuhan, China. From these 78 clinical features, 56 were used as input to a 6-layer fully connected deep neural network (Li et al., 2020). Clinical and CT scan data of 894 patients infected with COVID-19 and pneumonia, 328 patients not infected with COVID-19, and 299 patients suspected to be infected with COVID-19, were used to train a CNN and a 13-layer CNN respectively. Both predictions were then integrated using the Polyomial Linear Regression (PLR) algorithm (Ning et al., 2020). Vital signs, laboratory data and ECG results of 567 patients were utilized to train a Random Forest Classifier to predict near-term mortality (20-84 hours before death). 99 features were selected based on implicit non-zero Gini importance values to train the model (Parchure et al., 2022).

5. Conclusion and future work

This study is aimed at determining the risk factors associated with mortality of patients infected with SARS-Cov2 and analyzes the role of blood biomarkers and a significant combination of features that contributes to predicting the patient’s mortality. An ensemble model consisting of a voting classifier was proposed to predict the mortality of patients infected with COVID-19 based on clinical data which can be obtained with a blood test. Machine learning models proposed in this paper can be used for the prognosis of infected patients as there were limited ICU beds available during the COVID wave, and predictive decisions could’ve been made to prioritize patients to be treated with the limited amount of resources available. A Voting Classifier with ‘soft’ voting was used to build an ensemble model. Baseline classification models with better results were chosen as estimators for the voting classifier i.e. Random Forest Classifier, XBoost Classifier, Gradient Boosting Classifier, and Extra Tree Classifier. To achieve optimal hyperparameters of the selected estimators, hyperparameter tuning was done using the grid search CV method. Improvement in accuracy, precision, and F1-score was observed for the proposed ensemble model when compared to other baseline models. F1 score and AUC-ROC were considered to be critical in model evaluation, in contrast to a few other studies, which employed ROC-AUC and accuracy as a comparative evaluation criteria. The proposed model has been compared with the base estimators as well as other various state-of-the-art models on various evaluation criteria such as accuracy, ROC-AUC, F1 Score, precision and recall.

There are some limitations of this research. The data, mainly from hospitalized patients, was extracted around the time of patient admission. As a result, the predictive model may have some bias because these patients often have other serious complications, and thus a higher probability of death than the general public. This model does not apply to pregnant, breastfeeding, or underaged patients, as it was created specifically for patients already infected with COVID-19. Further prospective research would be required to confirm the predictive accuracy of our prediction model as this study was a retrospective study.

For future studies, multiple datasets could be used to improve the analysis and comparability of the models for mortality prediction. The application of other features such as chest X-rays and CT scans in conjunction with this technique may yield better results, but that was out of the scope of this research. And also, this study will be enhanced by including many factors or variables like demographic factors, geographical factors, and weather factors (temperature, humidity, wind speed, and rainfall). Future efforts can be made to improve mortality prediction by incorporating a computational intelligence algorithm, such as monarch butterfly optimization (MBO) (Wang et al., 2019), the earthworm optimization algorithm (EWA) (Wang et al., 2018), elephant herding optimization (EHO) (Li & Wang, 2022), moth search algorithm (MS) (Feng & Wang, 2018), slime mould algorithm (SMO) (Li et al., 2020), Hunger Games search algorithm (HGS) (Yang et al., 2021), Runge-Kutta (RUN) optimizer (Ahmadifar et al., 2021), colony predation algorithm (CPA) (Tu et al., 2021), or Harris hawks optimization (HHO) (Heidari et al., 2019). (Eqs. 1–4, Table 1).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmadifar, I., Heidari, A. A., Gandomi, A. H., Chu, X., & Chen, H. (2021). RUN beyond the metaphor: An efficient optimization algorithm based on Runge Kutta method. Expert Systems with Applications, 181, Article 115079.

Achtschel, David. (2021). Neurologic complications of COVID-19, dryad, dataset Available online at. 10.5061/dryad.74w3r7w.

An, C., Lin, H., Kim, D. W., Chang, J. H., Choi, Y. J., & Kim, S. W. (2020). Machine learning prediction for mortality of patients diagnosed with COVID-19: A nationwide Korean cohort study. Scientific Reports, 10(1), 1–11.

Bolourani, S., Brenner, M., Wang, P., McGinn, T., Hirsch, J. S., & Barnaby, D. Northwell COVID-19 Research Consortium. (2021). A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: Model development and validation. Journal of Medical Internet Research, 23(2), e24266.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society: Series B, 22(2), 302–306.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

Cheng, J., Sollev, J., Huiie, C., Yue, H., Vandal, N., Shanahan, J., & Wang, J. (2022). COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data. European Radiology, 1–11.
