Boosting Ensemble Machine Learning Approach for Covid-19 Death Prediction

Sri Lanka Journal of Social Sciences and Humanities
Volume 3 Issue 1, February 2023: 81-89
ISSN: 2773 692X (Online), 2773 6911 (Print)
Copyright: © 2023 The Author(s)
Published by the Faculty of Social Sciences and Languages, Sabaragamuwa University of Sri Lanka
Website: https://www.sab.ac.lk/sljssh
DOI: http://doi.org/10.4038/sljssh.v3i1.88

Kuhaneswaran Banujan[1,*] and Mohamed Ifham[2], and B.T.G.S. Kumara[3]

[1, 2, 3] Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, Sri Lanka.

Received: 15 September 2022, **Revised:** 30 November 2022, **Accepted:** 16 December 2022.
How to Cite this Article: Kuhaneswaran Banujan, Mohamed Ifham, and Kumara, B.T.G.S. (2023). Boosting ensemble machine learning approach for COVID-19 death prediction. *Sri Lanka Journal of Social Sciences and Humanities, 3*(1), 81-89.

Abstract

It is critical for physicians to correctly classify patients during a plague and determine who deserves minimal health assistance. Machine learning methods have been presented to reliably forecast the severity of COVID-19 disease. Previous research has often tested different machine learning algorithms and evaluated performance under different methods. It may be necessary to try several combinations of machine learning algorithms to discover the optimal prediction model to get the best results. This research aimed to train boosting ensemble algorithms and Artificial Neural Networks (ANN) and choose the model that best predicted how long patients would survive a Covid19 infection. The dataset for this study was obtained through kaggle.com. It contains blood samples from 4313 patients and is retrospectively evaluated to find relevant measures of overall mortality. Out of 48 parameters, only 16 selected parameters were considered using the information gain weight for each parameter. 5-fold cross-validation was employed on the training data set, and Receiver Operating Characteristic (ROC) curves were created to verify better the prediction algorithms' performance independent of the algorithm choice criteria. The models XGBoost, CatBoost, and LightBGM achieved an accuracy of 98%, AdaBoost 96%, and 93% for ANN, respectively, implying that ANN has lower accuracy than boosting approaches.

Index Terms- Artificial Neural Network, Boosting algorithms, Receiver operating characteristic, Covid19, Machine learning

INTRODUCTION

SARS-CoV-2 is a contagious infectious disease that began as a pandemic with a significant impact on global public health, resulting in mortality and severe health issues. The virus was first detected in the Chinese city of Wuhan in late 2019, where many individuals suffered pneumonia-like prodrome (Team, 2020). It has a variety of impacts on the body, including respiratory problems and multi-organ dysfunction, which leads to death in a short amount of time (Van Der Hoek et al., 2004). Till 25 February 2022, the Covid-19 pandemic has infected over 433 million people worldwide and killed over 5.94 million people, with the United States and Europe being the most impacted regions with rising fatality rates. This pandemic affects thousands of individuals worldwide, with hundreds of deaths expected daily.

Each day, thousands of new persons are claimed to be positive worldwide. Covid-19 is spread mainly through direct contact via droplets transmitted by coughing, sneezing, or speaking to an infected individual (Hu, Ge, Li, Jin, & Xiong, 2020). The investigation for transitory carriers through which the illness could have been transmitted to humans is ongoing; regardless of the initial source, the Covid-19 pandemic showed an extraordinary hematogenous spread. Its transmission is complicated because an individual might be exposed to the virus for days despite showing symptoms. Due to the apparent causes of its development and threat, practically all governments have announced total lockdowns in the affected areas. The fundamental motivation for these efforts was predicting models demonstrating that there would be fewer deaths without isolation procedures. Medical researchers worldwide are working to find an effective vaccination and treatment for the disease. Because there is currently no licensed therapy to treat the virus, authorities worldwide are working on preventative measures to halt its spread.

Without any complete medical remedy and the likelihood of new virus strains, the average worldwide mortality rate increases each day. Even though rigorous social distance and protective measures remain, the virus's global death and prevalence curves indicate no progress (Xia et al., 2019). A greater focus on early treatment options could be beneficial in lowering mortality rates. Patients in critical condition will need to be admitted to the Intensive Care Unit (ICU) immediately, necessitating the usage of ventilators. Doctors frequently cannot precisely forecast the diagnosis of Covid-19 victims until late in the virus's course. Moreover, the prognosis of Covid-19 can take unexpected turns, with a patient's health suddenly deteriorating to a catastrophic state after appearing stable (Cascella, Rajnik, Aleem, Dulebohn, & Di Napoli, 2022; Rajnik, Cascella, Cuomo, Dulebohn, & Di Napoli, 2021), even the most experienced clinicians the surprise.

* **Corresponding author:** Tel.: +94 (77) 748 8583; Email: bhakuha@gmail.com
https://orcid.org/0000-0002-0265-2198

This article is published under the Creative Commons CC-BY-ND License (http://creativecommons.org/licenses/by-nd/4.0/). This license permits use, distribution, and reproduction, commercial and non-commercial, provided that the original work is properly cited and is not changed in anyway.

Researchers are working on remedies to lessen the consequences of the Covid-19 pandemic as the globe grapples with the dilemma of the virus's deadly effects. Medical practices are encountering difficulty in providing proper diagnosis and treatment as the Covid-19 pandemic spreads worldwide, causing potentially fatal respiratory illnesses. Because of Covid-19's widespread distribution and severe consequences on humans, multiple study organisations have looked into the virus's epidemiological features, socioeconomic implications, and elements and parameters that aid the virus's transmission. Artificial Intelligence (AI) algorithms could be helpful to assist in enhancing patient prediction because they can recognise complicated patterns in massive datasets (Shilo, Rossman, & Segal, 2020; Yu, Beam, & Kohane, 2018), a capability that the human brain lacks.

Machine learning is a method in which computers analyse data and learn from the results. A neural network is created by simulating complicated algorithms that help systems analyse, evaluate, and understand data before applying the knowledge gained to resolve problems and make predictions. Over the last decade, machine learning algorithms have become an important research topic by addressing highly advanced and complicated real-world issues. Artificial Neural Networks (ANNs) have been widely employed to capture the uncertainties in unstructured datasets since they have been an effective tool for dealing with non-linear data (Hainaut, 2018). As a result, the usage of these ANN models in epidemiological estimates for linear, non-linear, and hybrid data has exploded in recent years.

Since the outbreak's start, experts and research organizations have been very interested in pattern analysis and forecasting of Covid-19's spread worldwide. Hospitalised individuals with Covid-19 are constantly in danger of death. Machine learning (ML) approaches could predict death in Covid-19 patients who are hospitalised. As a result, our research aimed to utilise machine learning algorithms for forecasting Covid-19 mortality using clinical treatment information gathered from patients at the time of their hospitalisation.

This study aims to use a machine learning-based approach to create mortality forecasting models for Covid-19 patients by considering their health checkups and classifying patients into low- and high-risk categories. These machine-learning models were so influential that they shifted the course of several countries' responses. Models for predicting morbidity and mortality and a data methodology have been critical tools for authorities.

The paper is structured as follows. Section 2 describes the Literature Review in the context of Covid-19 prediction. Section 3 clearly explains the proposed approach, while Section 4 displays the results obtained and discusses the results. Finally, Section 5 concludes the paper with future research directions.

LITERATURE REVIEW

Covid-19 death prediction design models have been the subject of many different investigations. Yadav et al. (Chou et al., 2021; Yadaw et al., 2020) used a large cohort to construct a very accurate Machine Learning based mortality forecasting model that considered the participant age and O_2 saturation during their medical contact and patient engagement. The most outstanding predicted criteria were age and minimal oxygen saturation during the interaction, which was considered in our findings. In Covid-19 locations around the United States, individuals aged 60 and over make up about 85% of all mortality (Bhatraju et al., 2020). The degree of hypoxia at diagnosis has been widely described as a substantial predictor of illness severity, particularly in severe respiratory infections. It has an excellent rationale for being a significant predictor component in the diagnostic workup of the Covid-19 pandemic (Duca, Piva, Focà, Latronico, & Rizzi, 2020; Grasselli et al., 2020).

Even though the research and testing datasets in this investigation were more extensive, the data obtained contained a detailed list of demographics, pathologies, biochemical testing, radiology, and omics data. Furthermore, despite having big datasets, the number of participants' deaths was low. Knight et al. evaluated an 8-item points system for in-hospital mortality owing to Covid-19 in a large prospective study (Knight et al., 2020). Age, gender, number of morbidities, breathing rates, O2 saturation, state of consciousness, urea level, and CRP. However, some potentially significant comorbidities such as hypertension, prior myocardial infarction, and stroke aren't included in data collection, resulting in low distinction for mortality. Furthermore, given the 32.2 percent mortality rate and older patient group, this model may perform differently in pediatric people or populations at reduced mortality risk.

According to LASSO and multivariate data assessment, increased age, coronary artery disease, the proportion of lymphocytes, procalcitonin, creatinine, CRP, and D-dimer could be significant risk factors for Covid19 deaths according to LASSO and multivariate data assessment estimation techniques. Using a robust prediction model, these characteristics could classify Covid cases into high - and low groups (Shang et al., 2020). Covid-19 fatality prediction models show significant heterogeneity. Lactate dehydrogenase and procalcitonin are among the top death-forecasting factors in the model that Zhao et al. (2020) created. The Covid-AID study revealed that renal failure at presentation, irrespective of chronicity, strongly affects in-hospital fatalities in hospitalised patients (Hajifathalian et al., 2020). Recent research has linked Prothrombin and CRP levels to COIVD severity and death (Bannaga et al., 2020; von Meijenfeldt et al., 2021). In this investigation, we found a link between lower O2 and greater lactate, which could indicate a higher level of metabolic activities (Li et al., 2020) in Covid19 patients, which is linked to death.

A systematic study and meta-analysis published in April 2020 found that severely affected Covid 19 patients have a much greater rate of hypertension, diabetes, cardiovascular disease, and respiratory disease than non-critical patients (Zheng et al., 2020). Furthermore, a comprehensive review and meta-analysis of risk factors for Covid 19 mortality, dyspnea, chest tightness, hemoptysis, expectoration, and exhaustion were the major significant clinical variables linked to an elevated risk of Covid-19 fatalities. Non-survivors had a significantly higher leukocyte count and a substantially lower lymphocyte count in this investigation (Yang et al., 2020).

Machine Learning was effectively used to forecast the need for ICU and mechanical ventilation in patients with Covid-19 (Patel et al., 2021). The most relevant parameters that determine the requirement for ICU include PCT, DD, CRP, respiratory rate, SpO_2, albumin, AST/SGOT, calcium, influenza-like symptoms, and ALT/SGPT, according to a random forest model. The most critical indicators to forecast the need for respiratory support were CRP, DD, PCT, SpO_2,

respiratory rate, creatinine, total protein, albumin, calcium, and age (Patel et al., 2021). The most relevant variables for predicting Covid severity in a comparable study were SpO2/FiO2, CRP, estimated glomerular filtration rate (eGFR), age, Charlson score, lymphocyte count, and PCT (Marcos et al., 2021). Leon et al. used a machine learning approach to divide Covid patients into high, moderate, and low mortality groups. This study linked higher and lowered AST, ALT, LDH, CRP, and neutrophil counts to a higher and lower death rate, respectively (Benito-León et al., 2021). Monocyte and lymphocyte percentages were found to be adversely linked with death.

Covid19 viruses are among the most common infections that attack the human respiratory system. The review aimed to identify the most common factors and methodologies to research the virus's spread. Infection is more likely in patients with chronic conditions such as diabetes, hypertension, diabetes, stroke, heart failure, and kidney failure and older people with weakened immune systems (Raghupathi, Ren, & Raghupathi, 2020). Infection risk may be increased in enclosed spaces with poor ventilation and airflow. Like other respiratory viruses such as influenza and rhinoviruses, the virus is thought to spread via droplets in the air from sneezing and coughing. Aerosol transmission can also happen when individuals are subjected to high aerosol concentrations in enclosed areas for an extended period (Team, 2020).

Several studies have outlined characteristics regarding quarantine infrastructure, laboratory testing facilities, and healthcare capability, all of which contribute to a state's pandemic preparation. These most significant and influential elements must be investigated as a pandemic remedy as soon as possible. The provision of open data sets matching various factors aids in speeding up studies and forming collaborations. Some research took into account environmental issues such as basic sanitation. Many researchers have considered Covid-19-related mortality and other demographic data (Soyiri & Reidpath, 2013). Comorbidities have been identified as a critical determinant in the frequency of Covid-19 patients in different research and studies (Guan et al., 2020).

Casualties may be misinterpreted as just Covid-19 mortality if comorbidities are not considered. Covid-19 mortality has been successfully predicted by experts from many universities worldwide. One such research, undertaken by Columbia University and the CDC (2020), employed "death" like an exponential curve and a susceptible-exposed-infectious-removed (SEIR), meta-population model to estimate social distance parameters. As can be observed from a primary evaluation of current models, various academics have sought to create statistical models of the Covid-19 epidemic since the pandemic. The scope, assumptions, projections, impacts of interventions, and their effects on health care are different (Kotwal, Yadav, Yadav, Kotwal, & Khune, 2020).

PROPOSED APPROACH

This research aimed to train several machine learning algorithms to predict the mortality rates of Covid19 infected patients. We chose the models that accurately predicted how long patients would survive a Covid19 infection. In addition, we determined which variables (e.g., epidemiological, demographic, clinical, laboratory, and mortality results) had the most significant impact on the model's accuracy. Figure 1 shows an abstract overview of the study.

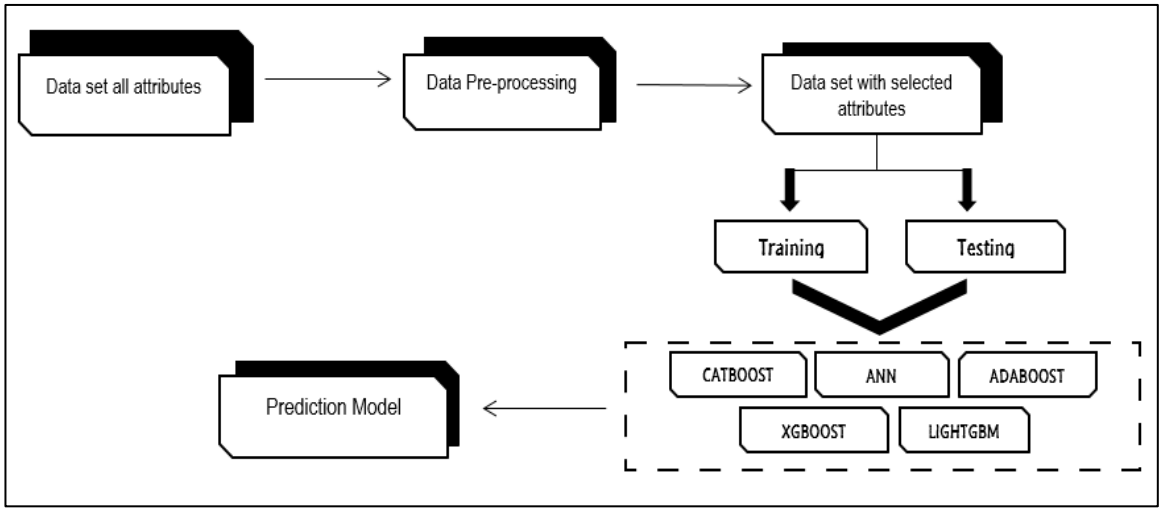

Figure 1: Methodological Approach

Dataset's Description

The dataset for this study was obtained through kaggle.com. It contains blood samples from 4313 patients and detailed laboratory reports of Covid19-affected patients and is retrospectively evaluated to find valid and relevant measures of overall mortality. Standard CSV files captured medical records, including clinical, demographic, laboratory, epidemiological, and mortality data. The dataset was published alongside the article by Ikemura et al. (2021), and the Montefiore Medical Center authorised the original work.

The data samples from patients over 18 who confirmed Covid-19 positive within 24 hours of admission were utilised in this study. We looked at 48 parameters in total and selected 16 parameters precisely considering the information gain weight of each parameter. By analysing each parameter's gain in the target attribute domain, information gain was utilised for feature selection. The criteria listed in Table 1 were used to diagnose and classify the severity of the patient.

Table 1: Detailed Description of the Attributes and Their Data Types

No	Attribute	Type
1	Age	Numeric
2	Charlson Comorbidity Index	Numeric
3	Number of ventilators used	Numeric
4	Diastolic BP	Numeric
5	Systolic blood pressure	Numeric
6	Creatine Level	Numeric
7	D-dimer	Numeric
8	Estimated glomerular filtration rate	Numeric
9	Ferritin Level	Numeric
10	Pulse Oximeter Level	Numeric
11	Respiratory rate	Numeric
12	Blood Urea Nitrogen (BUN)	Numeric
13	LDH test rate	Numeric
14	Procalcitonin test	Numeric
15	proBNP blood test	Numeric
16	Time from Covid positive to death in days	Numeric
17	Death (Outcome)	Boolean

Data Preprocessing

We investigated the feasibility of reliable mortality forecasting using clinical laboratory data based on the results of 4313 patients with Covid19 infection. Death prediction was performed utilising ANN and Boosting models to achieve this goal. All lab and clinical biomarkers were included in the ANN and Boosting models. These models are deemed optimum because all the needed biomarkers are available for prediction. The dataset was divided into training and testing to generalise the prediction model findings to an independent dataset. 5-fold cross-validation was employed on the training data set cases for algorithm training and tuning. Receiver operating characteristic (ROC) curves were created to verify better the prediction algorithms' performance independent of the algorithm choice criteria.

Data reduction is a critical step in the formation of machine learning models. When variables in a model associate with one another, they become duplicate variables. We could reduce calculation times in real-world clinical contexts by creating a model with several unique variables. Furthermore, data pre-processing allows models to handle data sparsity by using variables with more datasets. Again, by focusing on the top 16 most relevant factors, physicians may focus on requesting medical tests rather than collecting data on 48 factors.

After analysing the dataset's clinical aspect and utilising the Information gain process for feature selection, we identified the 16 most influencing factors to generate the machine learning models in this study to predict mortality. The influence of each parameter was initially prioritised based on the Information gain values in the best-performing models. Following that, we chose factors that were significant in these models. If the score of a variable differed between models, we picked variables based on clinical observations. We desired at least one distinct variable for each biological activity in clinical practice. We chose the variable with the fewest incomplete data points if more than one factor represented the same clinical physical activity.

The top 16 variables in the ANN and Boosting algorithmic models were: diastolic and systolic blood pressure, creatine level, age, Carlson comorbidity index, LDH test rate, pulse oximetry level, D-dimer, estimated glomerular filtration rate, Ferritin Level, Number of ventilators used, proBNP blood test, respiratory rate, BUN level, Procalcitonin test, and time_from_Covid_positive_to_death_in_days. We believed these factors were an excellent depiction of the biological processes impacted by SARS-CoV-2 infection. These factors are also simple to gather in clinical settings. They also cut down on the number of missing values. The levels of BUN and Ferritin, respectively, are indications of renal and cardiac function. We picked the D-dimer level over the fibrinogen level as our coagulation marker since it scored higher and had more data points.

Furthermore, we represented comorbidities by Charlson comorbidity scores. A high-ranking variable was glucose level. This was most likely attributable to the higher risk of death in diabetic individuals. We hypothesised that the Charlson comorbidity score was a much more comprehensive prognostic indicator than glucose level because it also influences diabetes.

Charlson Comorbidity measure is a value-weighted to determine the likelihood of death within a year after being admitted to the hospital for people with various comorbid conditions. A ventilator offers respiratory support to an individual who cannot breathe by transferring sufficient oxygen in and out of the lungs. In this study, the usage of ventilators for patients and the consequence of dealing with ventilators were considered. The diastolic value measures the pressure level in the arteries whenever the heart relaxes. That's when the heart receives oxygen and refills with blood. Anything beyond a 60 to 80 mmHg diastolic pressure level is considered unhealthy. The pressure applied as the heart beats and blood is discharged into the artery is systolic blood pressure. It is the highest number in blood pressure diagnostic testing reported as a percentage. The systolic blood pressure should be less than 120 mmHg for a healthy human.

Creatine is a biological waste material whose blood level indicates how efficiently the kidneys function. By supporting muscular development, creating contributes to the continuous energy source for active muscles. It's also small in the brain, heart, and other organs. A D-dimer protein fragment is formed when a blood clot melts in the body. When we are wounded, blood clotting is a crucial procedure that keeps us from bleeding excessively blood. When our wound has recovered, our body will usually eliminate the clot.

The estimated glomerular filtration rate (eGFR) is a test that determines our kidney disease phase by measuring our

kidney functions. It can be calculated by the medical team based on the findings of our blood creatinine test, age, body size, and gender. Ferritin is a blood protein that contains iron, and a trial will tell the physician how much iron is stored in the body. If a ferritin test depicts a lesser than average ferritin rate in the blood, the body's iron levels are insufficient and would have iron deficiency. The doctor may recommend a pulse oximeter if we have a symptom of breathing problems that cause lung or heart disease. It's an electrical instrument that measures RBC oxygen saturation. The respiratory rate is a clinical indicator showing how much air enters and leaves the lungs. A change in respiratory rate is often the first sign of degradation as the body attempts to keep oxygen supply to the tissues. The blood urea nitrogen (BUN) test provides essential information regarding the health of our kidneys. A BUN test measures the number of urea nitrogen in our blood. The urea nitrogen will persist in our blood if the kidneys aren't performing correctly. Urea levels in the normal blood range from 7 to 20 mg/dl. The kidneys may not function at total capacity if the BUN is more significant than 20 mg/dL. LDH is found in the heart, muscles, kidneys, pancreas, liver, brain, blood cells, and other organs and tissues.

The LDH testing is mainly used to determine the extent and location of tissue injury in the body. It's also sometimes used to track how far particular illnesses have progressed. The amount of procalcitonin in the blood is measured by a procalcitonin test. A high level may indicate a dangerous bacterial disease like sepsis. The body's extreme response to infection is known as sepsis. The NT-proBNP blood test detects cardiac arrest by measuring brain natriuretic peptides. If the doctor prescribes a BNP test, it is most likely experiencing heart failure. BNP levels may rise due to intrinsic cardiac dysfunction and other factors such as pulmonary or renal illness.

Feature Identification and Patient Selection

The indicated features for determining fatality in Covid-19 patients were defined in this phase. The first part determined the most critical clinical parameters through an extensive literature search in digital databases. A holistic features machine learning technique is then developed using medical reports from the patient's well-being, clinical symptoms, laboratory tests, and treatment classifications.

Patients under the age of 18 were not allowed to participate. These individuals should be considered as part of the pediatric investigation. The CLG registry database identified datasets from 4313 patients. Fifteen incomplete samples with excess missing data (more than 75%) were excluded from the study. The additional missing data were imputed using each variable's mean or mode. The data was verified for noise, aberrant values, errors, duplication, and nonsensical information. We approached the relevant physicians for alternate interpretations of data preparation. The total sample size utilised for this study was 4298 admitted patients over 18.

Model Implementation

Any form of processing conducted on raw data to ready it for future processing is called data pre-processing. In the mining process, this approach was utilised as a preliminary step. On the other hand, these strategies are currently being employed in machine learning models, AI-type models, and making judgments against each other. Furthermore, this method might be applied to a wide range of datasets.

We create 5 supervised learning models using Python, including four boosting algorithms: ANN, AdaBoost, CatBoost, XGBoost, and LightBGM. As part of the model implementation process, the following methods were used.

I. Initially, the PyCharm IDE was used to create a Python project. The machine learning requirements and external libraries were loaded into the project directory. The acquired dataset was then saved in the project directory.
II. The dataset was interpreted using the "Pandas" package.
III. Examine the dataset's layout and the number of rows and columns.
IV. Assessing for duplicate and missing values
V. Missing and duplicating values may cause certain final prediction model development inconsistencies. As a result, the missing values have been substituted with the mean value to make the prediction model unbias.
VI. After the data filtering and interpretation phases were completed, the data set was separated into 70% training and 30% testing groups.

The selected features were fed into supervised learning models such as ANN, XGBoost, AdaBoost, CatBoost, and LightBGM to determine the best mortality predictions based on the underlying assessment metrics. The trained models were tested with the testing data, and the classifiers with the optimal outcomes were chosen based on Accuracy, Precision, Recall, F1-Score.and Support. We successfully examine and contrast current and prior methodologies using these basic assessment measures.

Machine Learning Classification Models

Boosting Algorithms

Boosting Machine Learning is a more advanced and complex technique for solving complicated, data-driven real-world problems. Boosting algorithms are an ensemble learning strategy that uses several Machine Learning techniques to convert weak learners into strong learners, improving the model's accuracy. The ensemble approach integrates numerous learners to enhance the performance of Machine Learning algorithms. This learning method produces more efficient and accurate models than a single model. The boosting algorithm's primary premise is to create numerous weak learners and integrate predictions to construct a single strong rule. Basic Machine Learning techniques are used to generate these weak rules, which are then applied to various data distributions. These algorithms generate weak rules with each iteration. After numerous iterations, the weak learners are joined to create a strong learner that can forecast a more reliable conclusion. This study used four boosting algorithms (XGBoost, AdaBoost, CatBoost, and LightBGM) to predict Covid infection patients' mortality.

ANN

ANNs are artificial intelligence that simulates how the human brain processes a sequence of stimuli to produce an output. Three neurons make up an ANN: an input neuron, a hidden neuron, and an output neuron. Neurons are the building blocks of an ANN, best described as a weighted directed graph, as seen in Figure 2. Directed edges with weighted values reflect the link between outputs and inputs neurons.

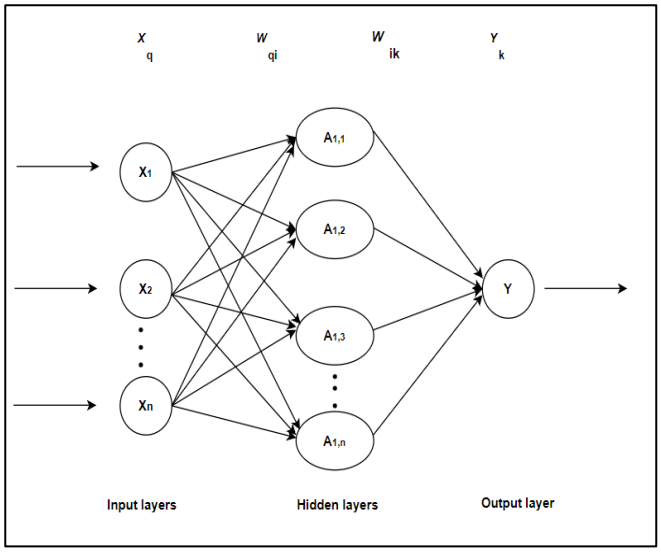

Figure 2: ANN Model

The ANN, which has a bias, computes the weighted sum of the inputs. Equation 1 uses a transfer function to express this calculation.

$$\sum_{i=1}^{n}(w_i * x_i) + b \quad (1)$$

The weighted total creates the output as input to an activation layer. Activation functions determine if a node should fire or not. Only those who have been fired have made it to the output layer. Depending on the task at hand, we can apply various activation functions.

The ANN was created using the python programming language. The hidden layer used the hyperbolic tangent activation function, and the output layer used the ReLu activation function with a cross-entropy error function. An input layer of 16 normalised variables, a hidden layer, and an output layer made up the network. The model batch trained over 70% of the samples, using scaled conjugate gradients as the optimisation strategy, and then evaluated the remaining 30%. The five-fold cross-validation approach was used to validate the model. 70% of all samples were assigned randomly to the training process, while the remaining 30% were assigned randomly to the testing stage, and the diagnostic ability was tested ten times.

RESULTS AND DISCUSSION

Evaluation Matrices

Evaluation matrices assess the efficiency of the research study's supervised machine learning algorithms. This usually entails utilising the dataset to train a model, then using the model to make forecasts on a data set derived from the testing data, and comparing the predictions to the predicted values in the remaining dataset. We used the specific dataset for each model to compare and evaluate whether the model was more effective for our research. Precision (equation 2), recall (equation 3), F1-score (equation 4), and accuracy (equation 5) are used to assess the efficiency of the proposed classification model for ANN, LightBGM, CatBoost, AdaBoost, and XGBoost.

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

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = \frac{2(Recall*Precision)}{(Recall+Precision)} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

Where;

- TP = Correctly predicted positive mortality rate
- TN= Correctly predicted negative mortality rate
- FP = Incorrectly predicted positive mortality rate
- FN= Incorrectly predicted negative mortality rate

Table 2 exhibits the evaluation results of precision (%), recall (%), f1- score (%), support, and accuracy for the ANN, XGBoost, AdaBoost, CatBoost, and LightBGM models.

Table 2: Individual Classifiers' Performance against Basic Performance Metrics

Classifier	Target Variable	Evaluation Metrics				Accuracy (%)
		Precision (%)	Recall (%)	F1- Score (%)	Support	
XGBoost	Survive	91.30	92.00	91.22	958	91.56
	Death	91.28	86.55	89.78	332	
LightGBM	Survive	91.22	92.63	92.56	958	90.48
	Death	91.25	87.77	89.14	332	
AdaBoost	Survive	89.38	91.80	90.17	958	89.59
	Death	87.42	82.88	84.47	332	
CatBoost	Survive	91.41	93.73	92.55	958	90.51
	Death	92.30	86.64	89.77	332	
ANN	Survive	83.25	75.82	79.74		86.78
	Death	83.29	75.87	79.26		

Confusion Metrics and Area Under Curve

The performance of ANN, CatBoost, XGBoost, AdaBoost, and LightBGM is compared using the basic evaluation metrics in this section. The evaluation results for the abovementioned methods corresponding to the dataset are shown in Table 3.

Table 3: Confusion Metrics and Area Under Curve Results for Individual Classifiers

Classifier	TN	FP	FN	TP	AUC
XGBoost	951	7	22	310	0.92
LightGBM	950	8	19	313	0.92
AdaBoost	938	20	36	296	0.89
CatBoost	954	4	22	310	0.92
ANN	958	27	52	253	0.86

A confusion matrix is a way of summarising the classification algorithm's efficiency. If the dataset has more than two classes or has an imbalanced number of observations in each class, classification accuracy alone can be deceptive. A confusion matrix and the standard assessment metrics for validating the model can be more beneficial. Figures 3 shows the confusion matrix results for (a) ANN, (b) XGBoost, (c) CatBoost, (d) AdaBoost, and (e) LightGBM, respectively.

We generated five machine-learning models using the top 16 influential variables and ranked them in terms of AUC. The LightGBM of each machine learning algorithm family has the best performance (AUC=0.97). With an AUC of 0.90, the ANN model was the least-performing independent model. The XGBoost (AUC=0.96) and CatBoost (AUC=0.96) had similar results, while the distributed AdaBoost model (AUC=0.94) came in 4th.

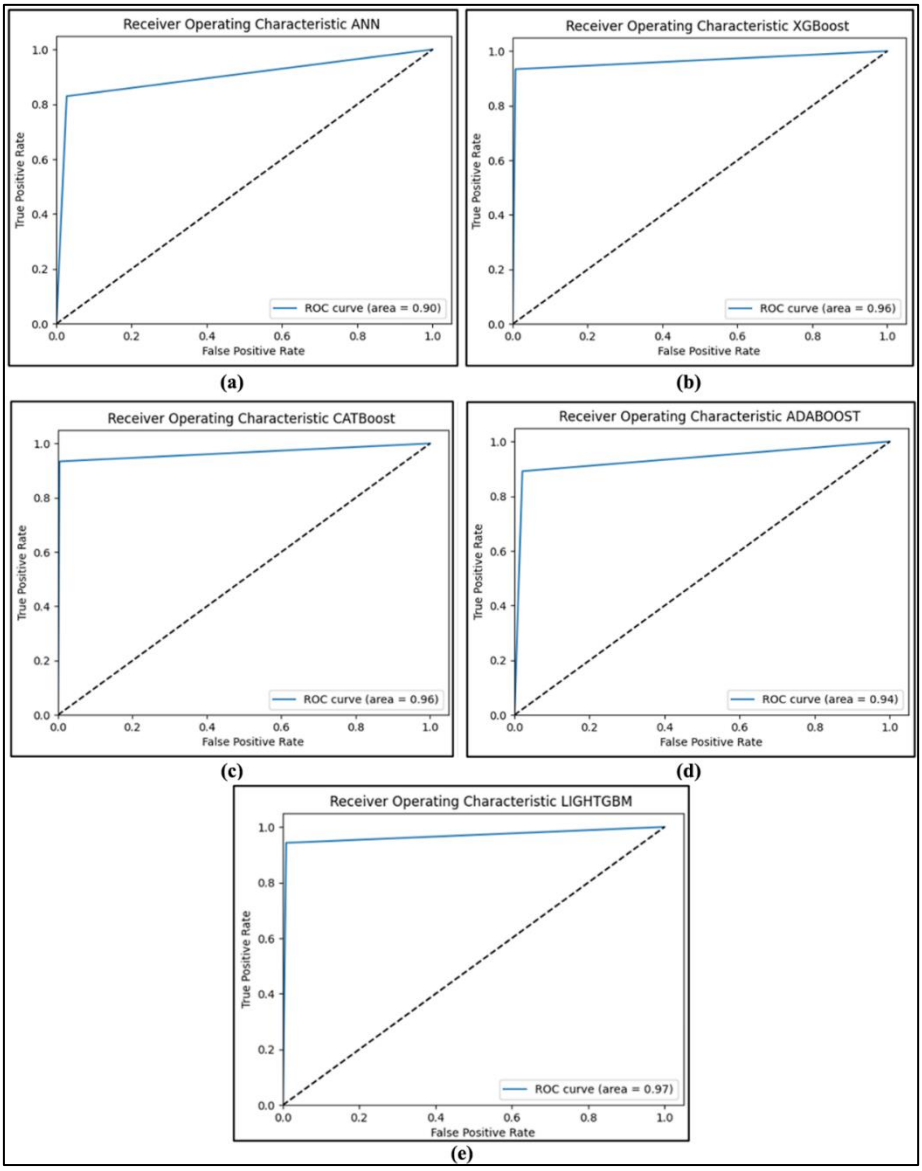

Figure. 3. ROC curve for (a) ANN, (b) XGBoost, (c) CatBoost, (d) AdaBoost and (e) LightGBM

Comparison of Performance

We effectively generated different machine learning models, evaluated their performance, and selected the best-performing models for predicting patients' odds of living after Covid 19 infection using Machine learning and clinical values (i.e., data collected early during a patient's enrollment in a hospital). Furthermore, our research shows that machine learning models based on only 16 clinical characteristics can predict survival. These models show outstanding responsiveness, specificity, and NPVs as well. As a result, ANN and Boosting models are effective, informative, and repeatable solutions.

According to our findings, the best models were LightGBM, CatBoost, and XGBoost (Shown in table 2). Their primary evaluation factors (i.e., Precision, Recall, F1- Score, and Support) indicated that they performed well. The ANN and AdaBoost models fared significantly worse than the LightGBM, CatBoost, and XGBoost models. The ANN model might have performed better if we had a larger data set. However, it took significantly longer to train than the other models. In interpreting tabular data, tree-based machine learning algorithms (e.g., LightGBM, CatBoost, XGBoost, and AdaBoost) are more efficient and probably more effective than neural network techniques. We chose the AUC as our model utility metric because it includes two critical clinical performance parameters that we were interested in: positive predictive value and False positive rate. We sought to explore which patients were most likely to die so we could intervene and rehabilitate as many as possible. On the other hand, the AUC considers model sensitivity and specificity while ignoring the impact of mortality occurrence on model performance. The model's performance depends on the prevalence of death; without such information, the model is useless.

Machine learning algorithms can improve e-medical record systems and determine patients' clinical information values. Our data pre-processing technique was successful based on the outcome of our machine learning models, and the particular 16 variables were sufficient to build high-performing models. This demonstrates that not all variables are required for computations and predictions. Physicians and hospitals should prioritise ordering medical checks as the first step in the patient review process (i.e., tests for the 16 variables shown in table 1). We cut down the number of variables we were required to evaluate and lowered the number of missing items in the dataset and the likelihood of imputation bias.

The goal of this machine learning model isn't just to forecast how long Covid-19 patients will live. These machine-learning models can be used to create models based on various clinical data for predicting multiple outcomes (e.g., forecasting models in which Covid infected patients require a ventilator). We believe our research will aid other researchers in implementing our Machine learning approach, speeding up the integration of artificially intelligent models into medical systems, and providing better medical treatment.

The proposed method is based on a performance analysis of numerous models for predicting the Mortality of Covid 19 infected patients. Because the target classes are closely related variables and have been rigorously reviewed in this comprehensive research, this type of argument has produced a consistent manner to provide a feasible option for machine learning algorithms.

The most fundamental performance metric is accuracy, simply the proportion of adequately predicted observations across all observations. As shown in Table II, our models XGBoost, CatBoost, and LightBGM achieved an accuracy of 98 percent, AdaBoost 96 percent, and 93 percent for ANN, respectively, implying that ANN has lower accuracy than boosting approaches and that the boosting algorithm, as a result of its superior accuracy, recall, and F1-Score value, outperformed ANN algorithms and has the best accuracy rate for a machine learning model. Therefore, the prediction model is built through the Boosting algorithms, a reliable mortality prediction.

Limitation of the Study

We acknowledge that our research had limitations. Patients with serious illnesses who needed admission to the hospital made up our sample. As a result, our results may not apply to all Covid-19 patients. The most critical variables in our study were systolic and diastolic BP. On the other hand, these characteristics could signal that people with severe diseases and hypotension are on the verge of dying. Our study did not take into account temporal characteristics. We didn't look at whether hypertension at arrival was a significant factor in patients who survived the first day of their stay. Furthermore, we did not investigate whether the significance of a variable decreases in populations that persist after the first 2 or 3 days. In the future study, we'd like to see if our algorithms can accurately predict death at various periods during admission.

In generalisation, more studies will be needed to see if our models apply to other institutions throughout Covid-19's waves. Because patient demographics fluctuate per institution, healthcare facilities must tailor their models to fit their specific patient groups. Models should be able to incorporate new data and adapt to the always-changing environment. We are constantly developing reinforcement learning approaches to keep our model up to date in real time.

This research was conducted in a retrospective manner. As a result, when patients were hospitalised, supervision could not improve the quality of data documentation. Furthermore, the study's data collection period coincided with the first pandemic outbreak, and health records were hurriedly logged as high patient loads. Due to a shortage of medical personnel, the medical system was compelled to prioritise patient care. As a result, many patients' medical profiles were incomplete and were sieved before the data evaluation step. The study's sample size was limited due to the considerations indicated above. Due to the pandemic state's healthcare and preventive regime limitations, no independent validation data was used. Furthermore, due to the increased missing values, limited laboratory resources, and inadequate health records caused by the pandemic, qualitative CRP, a characteristic linked to disease severity in multiple studies, was omitted from the analysis.

CONCLUSIONS

COVID-19, an infectious illness caused by the SARS-CoV-2. World-wide efforts are being made to establish a viable immunization and treatment for the illness. This study aims to train boosting ensemble algorithms and Artificial Neural Networks (ANN) and choose the model with the most accurate prediction of patients' Covid19 infection survival time. This study's dataset was collected from kaggle.com. It comprises blood samples from 4313 individuals and is analyzed retrospectively to identify pertinent metrics of

total mortality. Using the information gain weight associated with each parameter, only 16 parameters were examined out of a total of 48. The training data set was subjected to 5-fold cross-validation, and Receiver Operating Characteristic (ROC) curves were generated to validate the prediction algorithms' performance independent of the algorithm selection criteria. The models XGBoost, CatBoost, and LightBGM attained an accuracy of 98%, AdaBoost reached 96%, and ANN achieved 93%, indicating that boosting techniques are more accurate than ANN.

This model can be modified to estimate mortality for any kind of pandemic, such as covid, if the most influential characteristics of a new virus can be inputted with appropriate data. As future work, the authors intend to implement deep learning and deep ensemble techniques. Additionally, this study may be expanded to include other variables influencing covid 19 mortality.

REFERENCES

Bannaga, A. S., Tabuso, M., Farrugia, A., Chandrapalan, S., Somal, K., Lim, V. K., Mohamed, S., Nia, G. J., Mannath, J., & Wong, J. L. (2020). C-reactive protein and albumin association with mortality of hospitalised SARS-CoV-2 patients: A tertiary hospital experience. *Clinical Medicine, 20*(5), 463.

Benito-León, J., Del Castillo, M. D., Estirado, A., Ghosh, R., Dubey, S., & Serrano, J. I. (2021). Using Unsupervised Machine Learning to Identify Age-and Sex-Independent Severity Subgroups Among Patients with COVID-19: Observational Longitudinal Study. *Journal of Medical Internet Research, 23*(5), e25988.

Bhatraju, P. K., Ghassemieh, B. J., Nichols, M., Kim, R., Jerome, K. R., Nalla, A. K., Greninger, A. L., Pipavath, S., Wurfel, M. M., & Evans, L. (2020). Covid-19 in critically ill patients in the Seattle region—case series. *New England Journal of Medicine, 382*(21), 2012-2022.

Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S. C., & Di Napoli, R. (2022). Features, evaluation, and treatment of coronavirus (COVID-19). *Statpearls [internet]*.

Chou, E. H., Wang, C.-H., Hsieh, Y.-L., Namazi, B., Wolfshohl, J., Bhakta, T., Tsai, C.-L., Lien, W.-C., Sankaranarayanan, G., & Lee, C.-C. (2021). Clinical features of emergency department patients from early COVID-19 pandemic that predict SARS-CoV-2 infection: machine-learning approach. *Western Journal of Emergency Medicine, 22*(2), 244.

Duca, A., Piva, S., Focà, E., Latronico, N., & Rizzi, M. (2020). Calculated Decisions: Brescia-COVID Respiratory Severity Scale (BCRSS)/Algorithm. *Emergency medicine practice, 22*(5 Suppl), CD1-CD2.

Grasselli, G., Zangrillo, A., Zanella, A., Antonelli, M., Cabrini, L., Castelli, A., Cereda, D., Coluccello, A., Foti, G., & Fumagalli, R. (2020). Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy. *Jama, 323*(16), 1574-1581.

Guan, W.-j., Liang, W.-h., Zhao, Y., Liang, H.-r., Chen, Z.-s., Li, Y.-m., Liu, X.-q., Chen, R.-c., Tang, C.-l., & Wang, T. (2020). Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *European Respiratory Journal, 55*(5).

Hainaut, D. (2018). A neural-network analyser for mortality forecast. *ASTIN Bulletin: The Journal of the IAA, 48*(2), 481-508.

Hajifathalian, K., Sharaiha, R. Z., Kumar, S., Krisko, T., Skaf, D., Ang, B., Redd, W. D., Zhou, J. C., Hathorn, K. E., & McCarty, T. R. (2020). Development and external validation of a prediction risk model for short-term mortality among hospitalised US COVID-19 patients: A proposal for the COVID-AID risk tool. *PloS one, 15*(9), e0239536.

Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (2020). Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112*.

Ikemura, K., Bellin, E., Yagi, Y., Billett, H., Saada, M., Simone, K., Stahl, L., Szymanski, J., Goldstein, D., & Gil, M. R. (2021). Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study. *Journal of Medical Internet Research, 23*(2), e23458.

Knight, S. R., Ho, A., Pius, R., Buchan, I., Carson, G., Drake, T. M., Dunning, J., Fairfield, C. J., Gamble, C., & Green, C. A. (2020). Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *bmj, 370*.

Kotwal, A., Yadav, A. K., Yadav, J., Kotwal, J., & Khune, S. (2020). Predictive models of COVID-19 in India: a rapid review. *Medical journal armed forces India, 76*(4), 377-386.

Li, Z., Liu, G., Wang, L., Liang, Y., Zhou, Q., Wu, F., Yao, J., & Chen, B. (2020). From the insight of glucose metabolism disorder: oxygen therapy and blood glucose monitoring are crucial for quarantined COVID-19 patients. *Ecotoxicology and Environmental Safety, 197*, 110614.

Marcos, M., Belhassen-García, M., Sánchez-Puente, A., Sampedro-Gomez, J., Azibeiro, R., Dorado-Díaz, P.-I., Marcano-Millán, E., García-Vidal, C., Moreiro-Barroso, M.-T., & Cubino-Bóveda, N. (2021). Development of a severity of disease score and classification model by machine learning for hospitalised COVID-19 patients. *PloS one, 16*(4), e0240200.

Patel, D., Kher, V., Desai, B., Lei, X., Cen, S., Nanda, N., Gholamrezanezhad, A., Duddalwar, V., Varghese, B., & Oberai, A. A. (2021). Machine learning based predictors for COVID-19 disease severity. *Scientific reports, 11*(1), 1-7.

Raghupathi, V., Ren, J., & Raghupathi, W. (2020). Studying public perception about vaccination: A sentiment analysis of tweets. *International journal of environmental research and public health, 17*(10), 3464.

Rajnik, M., Cascella, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2021). *Features, evaluation, and treatment of coronavirus (COVID-19)*. Retrieved from

Shang, Y., Liu, T., Wei, Y., Li, J., Shao, L., Liu, M., Zhang, Y., Zhao, Z., Xu, H., & Peng, Z. (2020). Scoring systems for predicting mortality for severe patients with COVID-19. *EClinicalMedicine, 24*, 100426.

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: challenges and promises of big data in healthcare. *Nature medicine, 26*(1), 29-38.

Soyiri, I. N., & Reidpath, D. D. (2013). An overview of health forecasting. *Environmental health and preventive medicine, 18*(1), 1-9.

Team, E. (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. *China CDC weekly, 2*(8), 113.

Van Der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J., Wolthers, K. C., Wertheim-van Dillen, P. M., Kaandorp, J., Spaargaren, J., & Berkhout, B. (2004). Identification of a new human coronavirus. *Nature medicine, 10*(4), 368-373.

von Meijenfeldt, F. A., Havervall, S., Adelmeijer, J., Lundström, A., Rudberg, A. S., Magnusson, M., Mackman, N., Thalin, C., & Lisman, T. (2021). Prothrombotic changes in patients with COVID-19 are associated with disease severity and mortality. *Research and practice in thrombosis and haemostasis, 5*(1), 132-141.

Xia, J., Pan, S., Zhu, M., Cai, G., Yan, M., Su, Q., Yan, J., & Ning, G. (2019). A long short-term memory ensemble approach for improving the outcome prediction in intensive care unit. *Computational and mathematical methods in medicine, 2019*.

Yadaw, A. S., Li, Y.-c., Bose, S., Iyengar, R., Bunyavanich, S., & Pandey, G. (2020). Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *The Lancet Digital Health, 2*(10), e516-e525.

Yang, L., Jin, J., Luo, W., Gan, Y., Chen, B., & Li, W. (2020). Risk factors for predicting mortality of COVID-19 patients: A systematic review and meta-analysis. *PloS one, 15*(11), e0243124.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering, 2*(10), 719-731.

Zheng, Z., Peng, F., Xu, B., Zhao, J., Liu, H., Peng, J., Li, Q., Jiang, C., Zhou, Y., & Liu, S. (2020). Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *Journal of infection, 81*(2), e16-e25.