MACHİNE LEARNİNG MODELS FOR PREDİCTİNG SURVİVABİLİTY IN COVID-19 PATİENTS

[1]Ijegwa David Acheme* and [2]Olufunke Rebecca Vincent

[1]Department of Computer Science, Edo University Iyamho, Nigeria
[2]Department of Computer Science, Federal University of Agriculture Abeokuta, Ogun State, Nigeria

Emails: acheme.david@edouniversity.edu.ng; vincentor@funaab.edu.ng

ABSTRACT

Covid-19 is a disease currently ravaging the world, bringing unprecedented health and economic challenges to several nations. There are presently close to five million reported cases in over two hundred (200) countries with fatalities numbering over three hundred and thousand (300,000) persons. This study presents machine learning models for the prediction and visualization of the significant factors that determine the survivability of COVID-19 patients. This study develops prediction models using a decision tree, logistic regression, gradient boosting, and logistic regression algorithms to identify the significant factors and predict the survivability of COVID-19 patients. The results of the simulation showed that the logistic regression model had the lowest prediction accuracy. The other three showed over 95% correct accuracy and indicated that the essential factors in determining patients' survivability were underlying health conditions and age. The findings of this study agreed with the medical claims that patients with underlying health challenges and those advanced in age are liable to have complications; hence, providing a research-based credence to this belief. This proposed model thus serves as a decision support system for the management of COVID-19 patients, as well as predict a patients' chances of survival at the first presentation at the hospitals.

Keywords: *Machine learning, Logistic Regression, Survivability Factors, Corona Virus, Random forest, Decision trees.*

1. INTRODUCTION

The coronavirus disease called COVID-19 was officially reported in December 2019 in the city of Wuhan in the central Hubei province of the people's republic of China [1]. It was first reported as a pneumonia case of few clusters with the first patients being sellers at the Wuhan wet market. With increasing cases, the World Health Organization and the health authorities in China quickly established the cause of the disease as belonging to the family of coronaviruses. It was thus called a Novel Corona Virus (2019-nCov). The first reported fatality arising from this new disease was reported on the 11th January, which was a 61-year-old man who had contracted the virus at the Wuhan seafood market [2]. The disease rapidly spread across the world over a couple of few weeks, prompting the WHO to declare it a Public Health Emergency of International Concern on January 30th, 2020, and on February 11th, 2020, the WHO gave the disease the name COVID-19 [3].

Coronavirus, whose names are derived from the appearance of the outer fringe enveloping proteins resembling crown ('*corona'* in Latin), come from the group of RNA viruses [4]. They are found in birds and mammals and are known to cause infections to the upper respiratory system in humans. They were responsible for the Severe Acute Respiratory Syndrome (SARS) and the Middle-East Respiratory Syndrome (MERS) epidemic of 2003 and 2012. Covid-19 is the current outbreak from the family of coronaviruses, and it is ravaging almost all the nations of the World, bringing the world's economy to its knees in so many unprecedented ways. Without any known cure or vaccination, economic activities have been shut down in efforts to curtail the spread of this virus, the World Health Organization (WHO) reports that more than fifty percent (50%) of humanity is under a form of restriction from economic activities [5]. The catastrophe of this virus is in two phases: human mortality and economic redundancy. The International monetary fund (IMF) has predicted a global economic depression that is worse than that of 1930 and the crude oil mainstay of the global economy running on the negative price for the first time in the history of the world.

Besides the huge economic effects of this disease, the ever-increasing mortality remains a considerable concern. While The WHO reports a 4% mortality rate [6], this is highly debatable as it appears that several cases of fatalities are unreported [7]. Considering the highly infectious nature of this disease and its spread across substantial populations, the total number of deaths has already exceeded that of previous coronavirus cases and still counting. As on the morning of 17th May 2020, a total of over 4 million confirmed cases has been reported from 204 countries of the world; also, there are over 300,000 confirmed deaths across the globe, as reported by the WHO [6].

With the huge challenge posed by this disease, several research efforts are being sponsored across the world, especially on the genome sequence of the virus, which will ultimately lead to the development of a vaccine [8 – 10]. Efforts have also been reported in studying the economic effects of this disease [11]. Other researches have focused on the study and modeling of the disease spread patterns among populations and cities of the world aimed at better understanding and predicting infections and mortality rates [12]. The focus of this research effort, however, is the prediction of the survivability of infected persons to understand the factors responsible for the majority of fatalities. In the United States and the United Kingdom, the rates of deaths have been higher among the Black, Asian, and Minority Ethnic (BAME).

The application of machine learning models in medical research has been reported in several works, especially in the predictability of survival and prognosis of cancer [13]. Heart Diseases [14], Kidney Diseases [15], Parkinson's Disease [16]. Random forest classifiers, decision trees, and artificial neural networks (ANNs) specifically were among the earliest used techniques in medical research [17 - 19]. The most recent applications of machine learning methods have been in the detection and classification of tumors using CRT and X-Ray image data [20 -21], PubMed statistics reports over 2000 published research works on the detection, classification, and survival/prognosis detection of diseases in humans.

A survival prediction model for Pulmonary Arterial Hypertension (PAH) disease is presented in [22]. The study was aimed at studying and identifying the factors that determine survival in pulmonary arterial hypertension (PAH) disease, as understanding those factors will lead to better management of patients. The research utilized data retrieved from the US registry to evaluate early and long term pulmonary arterial hypertension disease. The data was analyzed to identify the factors responsible for one-year survivability. Hence the independent prognosticators were identified, leading to a weighted multivariable risk formula for use in the clinical management of patients.

[23] presented a machine learning model for the prediction and visualization of prognostic indicators in breast cancer patients to predict survivability. Dataset consisting of over eight thousand records covering the period of 1993 - 2016 were retrieved from the University of Malaya Medical Center in Kuala Lumpur Malaysia. The dataset consisted of twenty-three predictor variables and one dependent variable, “survival,” which represents the survival of the patient. Prediction models were built using decision trees, random forest, neural networks, logistic regression, and support vector machines. The models’ results showed close outcomes in terms of accuracy with decision trees giving the lowest accuracy of 79.8%, while random forest gave an accuracy of 82.7%. Furthermore, the model revealed the most correlated

variables hence the most important in determining survivability; these are the stage of cancer, size of the tumor, number of axillary lymph nodes removed, number of positive lymph nodes, types of primary treatment, and methods of diagnosis.

The study of [24] also presented a survivability model for breast cancer patients. The research utilized the SEER dataset covering about 30 years, containing a total of 433,272 records of breast cancer incidences. The data after preprocessing to remove redundancies and missing fields resulted in 202,932 records, which were classified into two groups of "survived" and "not survived." Machine learning algorithms were then applied to identify the dependent field from the sixteen (16) predictor fields. The results of the prediction of survivability reported were over 93% accurate.

[25] showed an approach for predicting survivability in malignancy. The main factor used for predicting survival time is the initially evolved tumor-incorporated clinical feature, which is a combination of tumor stage, tumor size, and age at diagnosis. The research utilized datasets from corresponding breast cancer, which were integrated using document-oriented graph databases. The applied machine learning methods of linear Support Vector Regression, Lasso regression, Kernel Ridge regression, K-neighborhood regression, and Decision Tree regression showed promising results in terms of accuracy of survival time prediction. [26] presented a multi-model ensemble technique for lung, stomach, and breast cancer prediction. The ensemble technique utilized several deep learning-based classifiers for predicting cancer occurrence. [27] used clinical data of patients of the Iranian Center for Breast Cancer (ICBC) from 1997 to 2008. The dataset with 1189 records, 22 predictor variables, and one outcome variable. They implemented three machine learning models for prediction of cancer in the patients; these are; Decision Trees, Support Vector Machine (SVM), and Artificial Neural Network (ANN). The research objective was to compare the performance of these three well-known algorithms by sensitivity, specificity, and accuracy analysis. Comprehensive reviews of several machine learning techniques that have been applied to disease prediction and survivability are found in [28 – 29]. Other researches that have also reported the survivability prediction in known diseases using machine learning methods are found in [12, 30 – 35].

This research deploys data science techniques using Machine Learning classification algorithms trained by existing clinical data of COVID-19 cases to predict the survivability of patients, thereby leading to a better understanding of the factors most responsible for fatalities. Machine learning which is a branch of artificial intelligence utilizes tools in statistics and probabilistic optimizations to allow computers learn from data and hence able to detect patterns that are hard to discern from noisy, complex and large datasets, this capability of machine learning models have positioned its applications suitable for medical research especially in applications which depends on complex proteomic and genomic measurements.

The paper is organized as follows: Section 1.1 presents related machine learning survivability models deployed to study other diseases. In Section 2.0, the procedure for data collection, wrangling, and pre-possessing and feature selection are presented. Sections 3 presents the results. In section 4 we present a discussion and conlusion in section 5.

2. 0 MATERIALS AND METHOD

The duration of time that a patient had COVID-19 virus is essential to his survivability of the virus. This study presents a framework for the survival analysis of the COVID-19 Pandemic. In this case, it is crucial to know the population of the COVID-19 population that would be expected to survive the Pandemic and at what rate. For the patients who are unable to survive the virus, it is essential to note the rate of death and what could be another underlying ailment. The particular circumstances and characteristics increase or decrease in the probability of survival are also of interest.

This study utilizes the dataset of COVID-19 cases in Nigeria as a case study, which is daily tallied by the Nigerian Centre for disease control NCDC. The study follows the well-known data science research methodology, as proposed by [36], which is illustrated in Figure 1. Figure 1 presents a step level of the survivability analysis. The machine learning models used in this study are decision tree, random forest, logistic regression and gradient boosting machine learning classifiers have been used, while the Area under the ROC curve and *F*1 measure and other established evaluation metrics were used for evaluation as it applies to binary classification problems

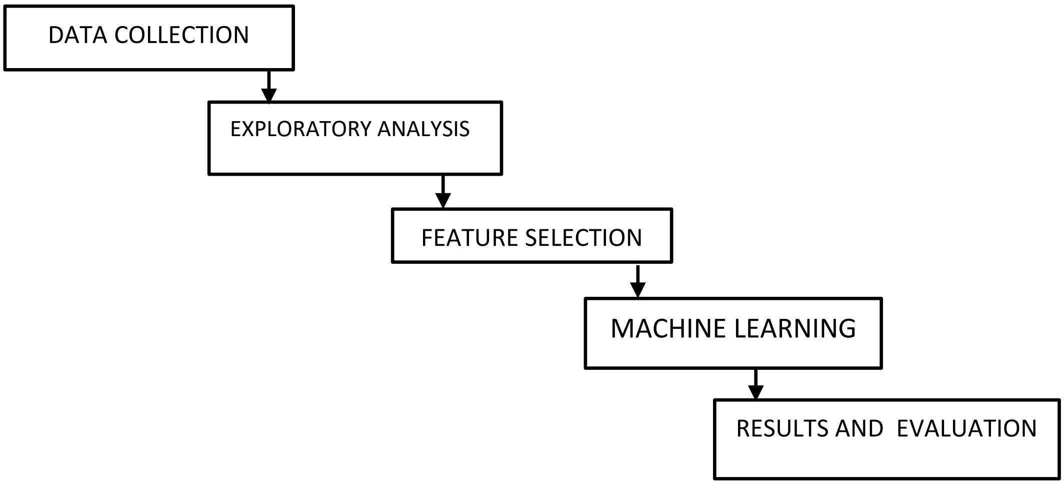

Figure 1. A Data Science Research Methodology

Data collected consisted of the fields presented in Table 1. An exploratory data analysis was then carried out to discover hidden patterns and gain further insights from the data leading to the removal of fields that were considered not very relevant to the prediction of survivability in the feature's selection. With the data cleaned and relevant features selected, the data was then spilled in a 70:30 ratio for training the chosen machine learning algorithm and testing the model, respectively. The results of the model were then evaluated using standard metrics of ROC AUC curve, F1 Measure, and log loss.

2.1 A Prediction of Survivability of the COVID-19 Patients using Machine Learning

The proposed COVID-19 survivability model comprises of following phases; Data collection, data pre-preprocessing, feature selection, building machine learning models, and comparative analysis of the models. Figure 2 describes the stages.

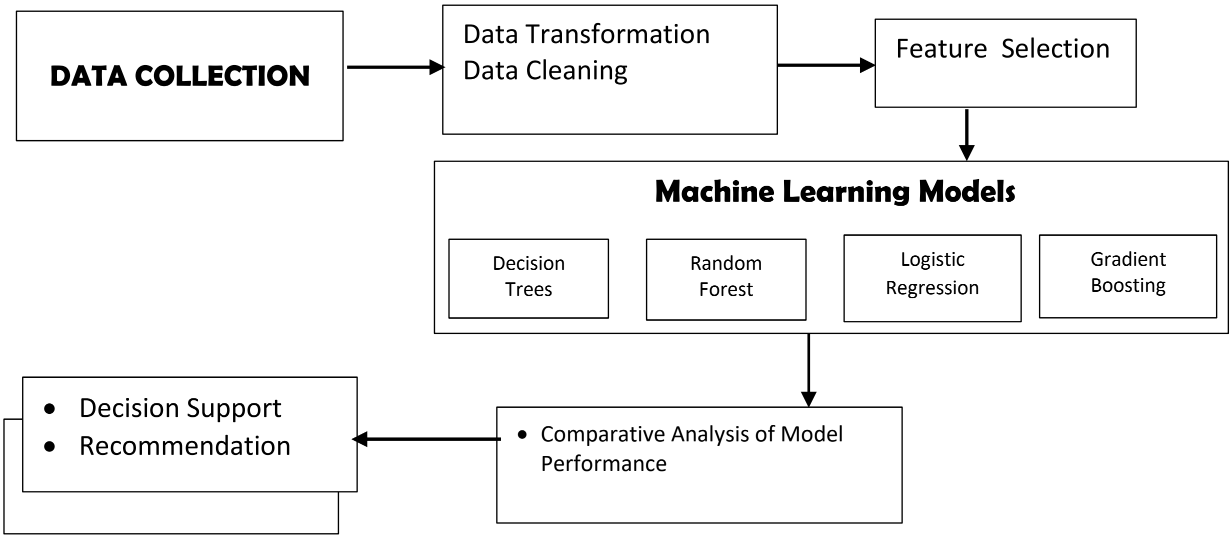

Figure 2. The proposed Covid-19 survivability architecture

From Figure 2, datasets containing records of COVID-19 cases in Nigeria are collected from the Nigerian Center for Disease control. The data is pre-processed and cleaned; the next exploratory data analysis is carried out to gain initial insights into the distributions of the variables. Four machine learning models are then built for comparative analysis and decision support.

2.1.1 Data Collection

The dataset after cleaning consisted of 1400 multivariate instances with attributes related to patient's age, marital status, race, occupation, gender, education level, employment status, overseas' travel history, other health conditions with the target variable being the survival status after one month Table 1 presents the summary of the variables selected from the data set. Table two shows the analysis and source of the dataset.

Table 1. Description of selected variables in the Nigerian COVID-19 Dataset

NAME	DESCRIPTION	VALUE(S)
Patient's Age	Age	Age
Marital Status	Patients' marital status	Yes or No
Race	Ethnicity	African, European, American, Asian
Occupation	Employment Status	Full employment, self-employment or students
Gender	Gender	Male or female
Level of Education	Extent of education	Not educated, educated up to Secondary School, educated up to University level
Overseas Travel History	Recent travel history to other countries in the past three Months	Yes or No
Other Health Conditions	Underlying health conditions	Diabetes, hypertension, cancer, etc
Status after one month	Survivability after one month of admission	Recovered or died

Table 2. An Analysis of Covid-19 Cases in Nigeria

DATE	WEEK	CASES	DISCHARGE	DEATH	AGE RANGE	SEX	UNDERLINING DISEASES	SOURCE
23rd -29th February	1	1	-	-	44	M	-	www.covid19.ncdc.gov.ng 29th February
1st -7th March	2	1	-	-	44	-	-	www.covid19.ncdc.gov.ng 7th March
8th -14th March	3	2	-	-	31-50	M	-	www.covid19.ncdc.gov.ng 14th March
15th-21nd March	4	25	2	-	35-60	M-70% F-30%	-	www.covid19.ncdc.gov.ng 21st March
22rd - 28st March	5	97	3	1	31-60	M-70% F-30%	Cardiac Arrest, Diabetes	www.covid19.ncdc.gov.ng 28th March
29st -4th April	6	214	25	4	31-50	M-70% F-30%	Immunodeficiency	www.covid19.ncdc.gov.ng 4th April
5th -11th April	7	318	70	10	31-60	M-73% F-27%	Hypertension	www.covid19.ncdc.gov.ng 11th April
12H -18th April	8	342	166	19	31-40	M-71% F-29%	Diabetes	www.covid19.ncdc.gov.ng 18th April
19th -25th April	9	1182	222	35	31-40	M-66% F-34%	Immunodeficiency	www.covid19.ncdc.gov.ng 25th April
26th -2nd May	10	2388	385	85	31-70	M-66% F-34%	Diabetes, Immunodeficiency	www.covid19.ncdc.gov.ng 2nd May
3rd -9th May	11	4151	778	143	31-70	M-66% F-34%	Immunodeficiency Pregnancy ,Diabetes	www.covid19.ncdc.gov.ng 9th May
10th -16th May	12	5959	1594	182	30-70	M-66% F-34%	Immunodeficiency ,Diabetes, Cancer	www.covid19.ncdc.gov.ng 16th May

2.1.2 Data Pre-processing

Data preprocessing is an iterative process for the transformation of the raw data into understandable and usable forms. Raw datasets are usually characterized by incompleteness, inconsistencies, lacking in behavior, and trends while containing errors [37]. The preprocessing is essential to handle the missing values and address inconsistencies. In this work the data gathering was carried out to avoid out-of-range values, impossible data combinations such as (Sex: Male, Pregnant: Yes) were handled, missing values and redundancies were also treated during the data preprocessing stage resulting in a more reliable and relevant dataset fit for knowledge discovery.

Transforming data into suitable formats for a particular machine learning problem is an essential consideration at the beginning of the project. The presence of irrelevant, redundant information, noisy and unreliable data significantly affects the model outcomes and knowledge discovery, making the training phase more difficult. The Data preparation and filtering steps take the most amounts of time spent on an ML project but worth it. The steps involved include; cleaning, Instance selection, normalization, transformation, feature extraction, and selection, the product of data preprocessing is the training set.

2.1.3 Feature Selection

Feature selection is among the essential steps in a machine learning project, and this is also referred to as variable and attribute selection since the interest is in the most critical attributes that influence the predicted variable, a good selection of features ensures; simplified models enhancing more natural interpretations by researchers and users, shorter training time saving computational resources; the avoidance of the curse of dimensionality, and the avoidance of overfitting [38]. Since this process involves the reduction of the number of input variables for the development of the model, it will lead to a reduction in the computational cost of the model as well as increase the model's performance. Statistical based feature selection method was employed in this work which involved the evaluation of the relationship between the target variable and the input variables and selecting the variables with the strongest correlation. The summary of the selected features is presented in Table 1.

2.1.4 The Machıne Learnıng Models

In machine learning (ML), Artificial Intelligence is applied through different statistical, probabilistic, and tools for optimization, which learns from patterns in training data to classify new data presented after training [39]. Machine learning techniques have been applied to statistical problems for analysis and interpretation of data. However, ML extends statistical methods by the usage of programming constructs such as Boolean logic, conditional statements if…else, and conditional probabilities for optimization, classification, and clustering problems. The foundation of Machine learning is firmly rooted in statistics and probability. Still, it offers more robust results as it allows inferences and decisions to be drawn from models that may not be possible with conventional techniques [40 – 41]. Statistical methods for example used in multivariate regression or correlation analysis assumes variable independence as such a strict statistical model with build linear combinations of such variables, in this kinds of scenario, statistical models are limited by non-linear, inter-dependent and conditional variables characteristic of most biological systems, in this kinds of situation, ML models offer better results [42]. The success of a good machine learning model depends on the understanding of the problem and the data used, understanding the assumptions and limitations of the chosen algorithms as the best models are dependent on the quality of training dataset [43]. Other problems are classified under the dimensionality of variables, overtraining, and overfitting of models [44].

(i) Decision Tree

Decision Tree classifiers are among well-known supervised learning algorithms. They are useful in solving regression and problems involving the classification of categorical variables. Decision trees create a training model that is used to predict the category or class of the dependent variable using a set of decision rules, as implemented in work, the decision tree proceeds from the root comparing values of the root attribute with the value of the new record presented to it to create a decision branch based on the comparison [45]. This research implements a categorical DT because of the nature of the target variable—the decision tree algorithm 1. Algorithm 1 represents a decision tree algorithm for the survivability of COVID-19 patients.

```
Algorithm 1: A decision Tree for Survivability of COVID-19 Patients
Input: COVID-19 Preprocessed Data Set
Output: Survivability (Yes/No)
```

Step 1: Record the patients' cases with COVID-19

Step 2: Start treatment and record the changes to calculate the Entropy (H) and Information Gain (IG) on the daily treatment of attribute S

Step 3: Select the attribute with the smallest entropy or highest information gain

Step 4: Split S to produce a subset of the data

Step 5: Continue iteration on each sub-set utilizing only unused attributes.

The entropy E(S) measures the randomness of the information of the medical changes in the patients, and it is defined by

$$E(S) = \sum_{i=1}^{c} -P_i log_2 P_i \ . \qquad (1)$$

In Eq. (1), S represents the current state of the patient, P_i is the probability of survival for any even i of state S. The information gain is computed as

$$Entropy(B) = \sum_{j=1}^{K} entropy(j, after). \qquad (2)$$

Eq. (2) is an expression of the surviving patients. In Eq. (2), B is the dataset before splitting, K is the number of subsets generated, and (j, after) is the jth subset after splitting.

(ii) Random Forest

Random forests build on simple decision trees, hence comprise of several numbers of separate decision trees operating as an ensemble system. In a random forest model, each tree produces a prediction for a class, the class with the majority of predictions; therefore, it becomes the final predicted value [46]. Random forests seek to deploy the power in numbers as very large units of decision trees which are uncorrelated but operating in a random forest to produce better results than the individual constituent tree. The total essential features in a random forest, thus, is the average of all the trees, such that

$$RFfi_i = \frac{\sum_{j \in alltree} normfi_{ij}}{T} . \qquad (3)$$

In Eq. (3), $RFfi_i$ is the importance of the feature, $normfi$ sub(ij) is the normalized importance i in tree j, and T is the total number of trees.

(iii) Logistic Regression

Logistic Regression (LR) belongs to the class of Generalized Linear Model (GLM) algorithms. Proposed in 1972 by Nelder and Wedderburn to provide a means of using linear regression to the problems which were not directly suited for application of linear regression. It is a classification algorithm widely used for building predictive models that utilize probabilities. It can be seen as a linear regression model with an associated cost function called the sigmoid or logistic function. This function maps predicted class values to the probability values between 0 and 1. The generalized equation is given in Eqn. 4

$$g\big(E(y)\big) = \alpha + \beta x1 + \gamma x2 \tag{4}$$

Where $g()$is the link function, $E(y)$ is the expectation of the predicted variable, and $\alpha + \beta x1 + \gamma x2$ are the predictors.

(iv) Gradient Boosting

Gradient boosting algorithms are machine learning techniques for classification and prediction problems. Gradient boosting works as an ensemble and optimization of several weaker models, such as decision trees. This classifier comprises three elements; A loss function which is optimized, A more inadequate learner such as decision tree to make predictions, and an additive function for adding up of weak learners to minimize the loss function.

3.0 COMPARATIVE ANALYSIS AND RESULTS

The model development involved the use of the entire dataset comprising of one thousand four hundred (n = 1400) records, which had eight (8) predictors of the survival rate variable. The dataset was split in the ratio 70:30 for training and testing, respectively. The four chosen models were built using IBM Watson studio, and each was evaluated with its accuracy, sensitivity, precision, F1 score, log loss, the receiver operating characteristic curve (AUC) and recall curve, and finally.

The decision tree was implemented utilizing the entire dataset. It processed the input data and yielded the tree with the optimal result with an accuracy of 95% correct prediction. The node of the DT signified the essential variable; these are followed by decision nodes that had percentages of classification. Figure 7(a) shows the feature importance of the decision tree classifier. In building the Random Forest (RF) model, 70% of the dataset was utilized for training. The RF model comprised of independent trees with the default number of trees set to (ntree = 500) to assess the model accuracy, the final prediction using the testing dataset (30%) yielding over 96% correct prediction.

Next was the logistic regression model (LR), this is a gaussian distribution with odds ratio, where the odds of the predicted variable (survivability) was modeled as a linear combination of all the predictor variables. The LR is useful in predicting binary depended on variables, in this case, the survivability, which is replaced in the dataset with 1 for alive and 0 for death. The LR model reported the least accuracy with the testing dataset. The gradient boosting classifier, which is an ensemble of random forest classifies, reported the highest accuracy. In this work, the model was built by converting the testing and training data into a matrix as xgboost for evaluation. The gradient boosting algorithm appeared to be the most suitable model for the prediction of survivability in COVID-19 patients.

The four machine learning models were built, trained, and evaluated using IBM Watson studio’s AutoAI tool on the IBM cloud. The complete dataset comprising of eight predictor variables and one target variable were used to build four machine learning models. For the evaluation of the model, the average precision, the area under ROC curve, precision, recall, F1 measure, normalized Gini coefficient, and log loss were the metrics used. Table 2 present the summary of COVID-19 cases in Nigeria. Tables 3 and 4 show a comparison of the evaluation metrics for the four chosen classifiers.

Exploratory Data Analysis is the process of initial exploration and investigation of the dataset to gain initial insights. In these ways, patterns and anomalies can be discovered. The results are presented as summary statistics and graphical representations Figures 3 to 5.

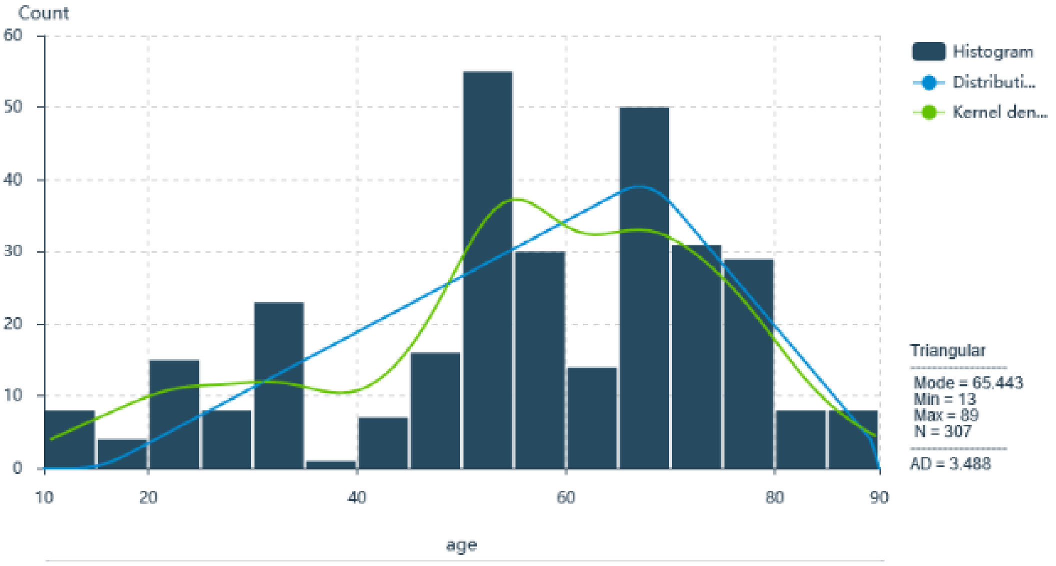

Figure 3. The Distribution of the variable Age of the Dataset

The age distribution shown in figure 3 reveals the age bracket of the most infected cases were between 50 – 55 and 60 – 70. While the minimum reported age was 15, and the maximum reported age was 89, indicating that the reported cases where well spread across the different ages in the population.

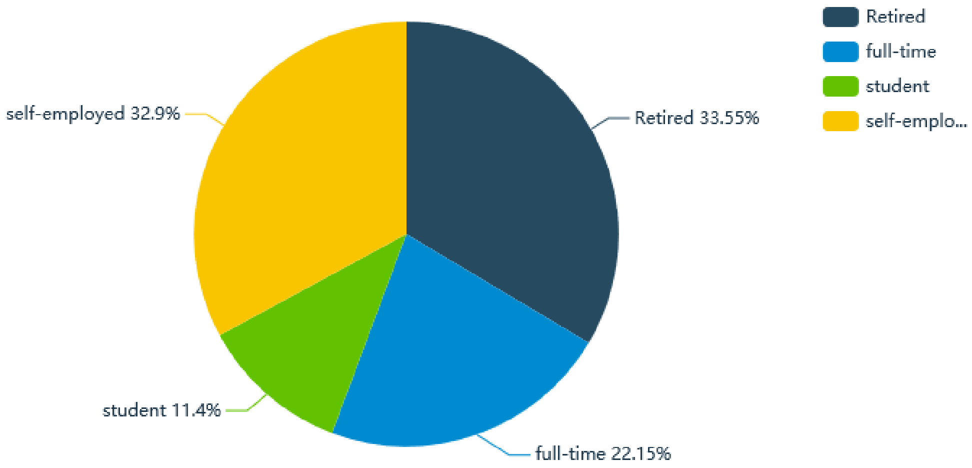

Figure 4(a). Frequency Distribution of variables occupation

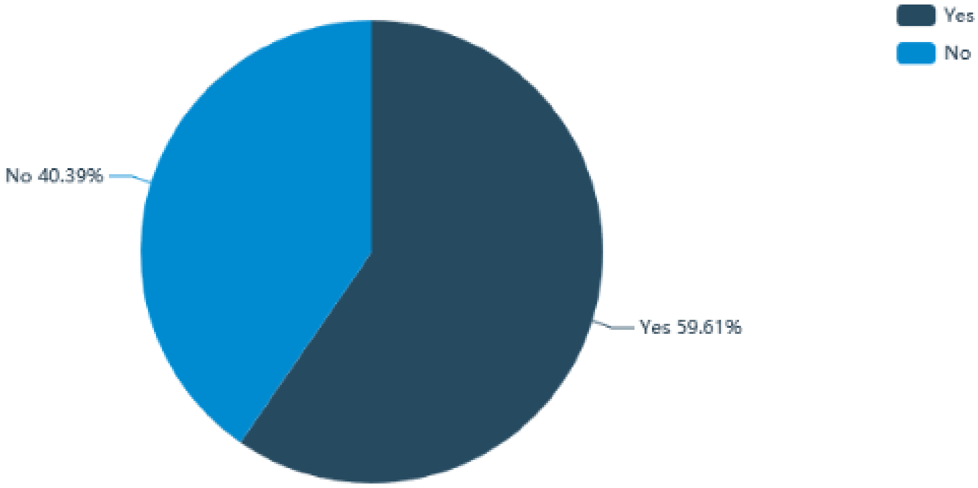

Figure 4(b). Frequency Distribution of variable Overseas Travel History

Further exploration of the data revealed about 33% of the patients admitted were business owners and self-employed, about 33% were retired from active service, the student population made up about 11%, and the fully employed were about 22%. Furthermore, in figure 4 (B), 59.6% of the reported cases had a travel history in the last three months, while 40.39% do not.

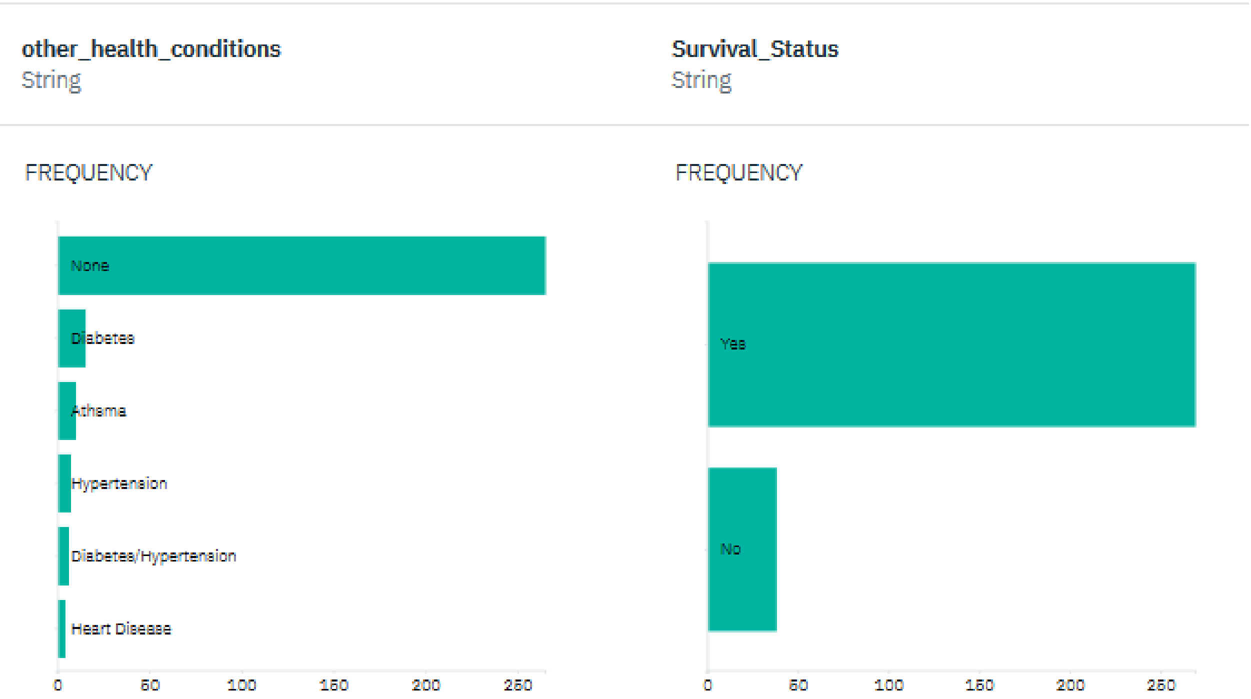

Figure 5. Frequency Distribution of variables Other Health Conditions and Survivability

Figure 5 is the frequency distribution of the patients with underlying health conditions and the total number of reported survivors after admission for one month. Figure 5 (A) reveals that 86.32% of the total cases had no known health conditions, 4.89% suffered from diabetes, 2.28% suffered from hypertension, 3.26% suffered from Asthma, 1.95% suffered from diabetes and hypertension. In contrast, about 1.3 suffered from other heart diseases. Figure 5(b) show that about 87% of admitted cases survived and were discharged within one month of admission, while about 13% of the cases were fatal.

3.1 Evaluation Metrics

The results of the decision tree, random forest, and gradient boosting classifiers showed over 95% prediction accuracy, while logistic regression showed an accuracy of 78.6%. See Table 4. Furthermore, a comparison of the feature importance of each algorithm is investigated, as presented in Figure 7, revealing that survivability of COVID-19 patients depended mostly on underlying health issues followed by age and occupation.

The performances of the models were evaluated with the AUC-ROC, F_1 Score, precision, and recall. These are summarized in Tables 3 and 4. The AUC-ROC, which is one of the most commonly used and reliable metrics, represents the extent or measure of separability, and it reveals the degree to which the models are capable of identifying classes. Higher values of AUC indicates better predictive accuracy. The ROC is plotted with the true positive (TPR) on the y-axis against false positives rates (FPR) and the x-axis. These values are estimated by Eqs. (5-7).

$$TPR/Recall/Sensitivity \ = \frac{TP}{FN} \tag{5}$$

$$Specificity = \frac{TP}{TN + FP} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

TP represents true positives, while FN is false negatives. TN represents true negatives, and FP denotes false positive. Figure 6 is the AUC-ROC curve for the gradient boosting classifier.

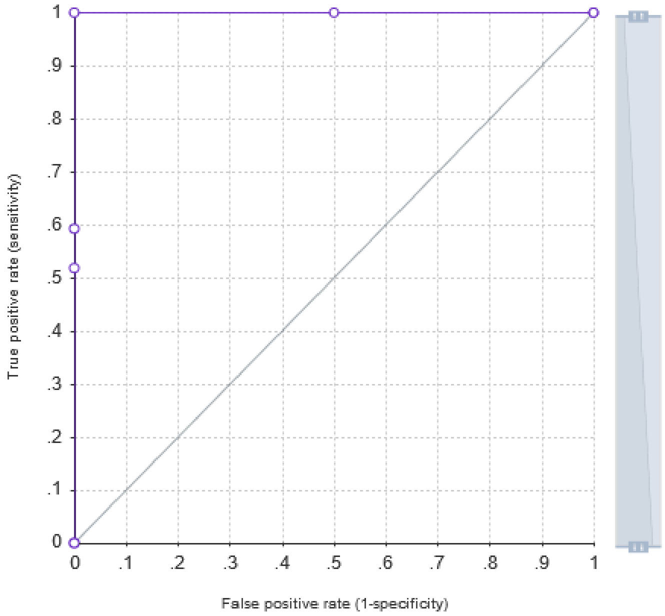

Figure 6. AUC-ROC Curve for the Gradient boosting algorithm

Figure 6 (AUC-ROC) for the gradient boosting classifier shows the value of almost 1. This reveals a very high measure of separation among the classes. Good models have AUC values close to 1 while poor models have AUC close to the 0 which means it has the worst measure of separability among classes and cannot be relied upon, as a matter of fact, it means a prediction of the opposite value. AUC-ROC values of 0.5 indicate a no class separation capacity in the model.

Table 3. Comparison of Algorithm performance

Algorithm	Performance (% Accuracy)
Decision Tree Classifier	95.5
Random Forest	96.4
Logistic Regression	78.6
Gradient Boosting Algorithm	99.3

Table 4. Comparison of performance metrics of the four classifiers

Algorithm	Avg Precision	F_1	Log Loss	Normalized Gini Coefficient	Precision	Recall	ROC AUC
Decision Tree Classifier	0.732	0.822	1.510	0.776	0.861	0.795	0.887
Random Forest	0.924	0.826	0.321	0.746	1.00	0.710	0.965
Logistic Regression	0.573	0.512	0.439	0.762	0.356	0.917	0.892
Gradient Boosting Algorithm	0.952	0.970	0.071	0.940	1.00	0.944	0.973

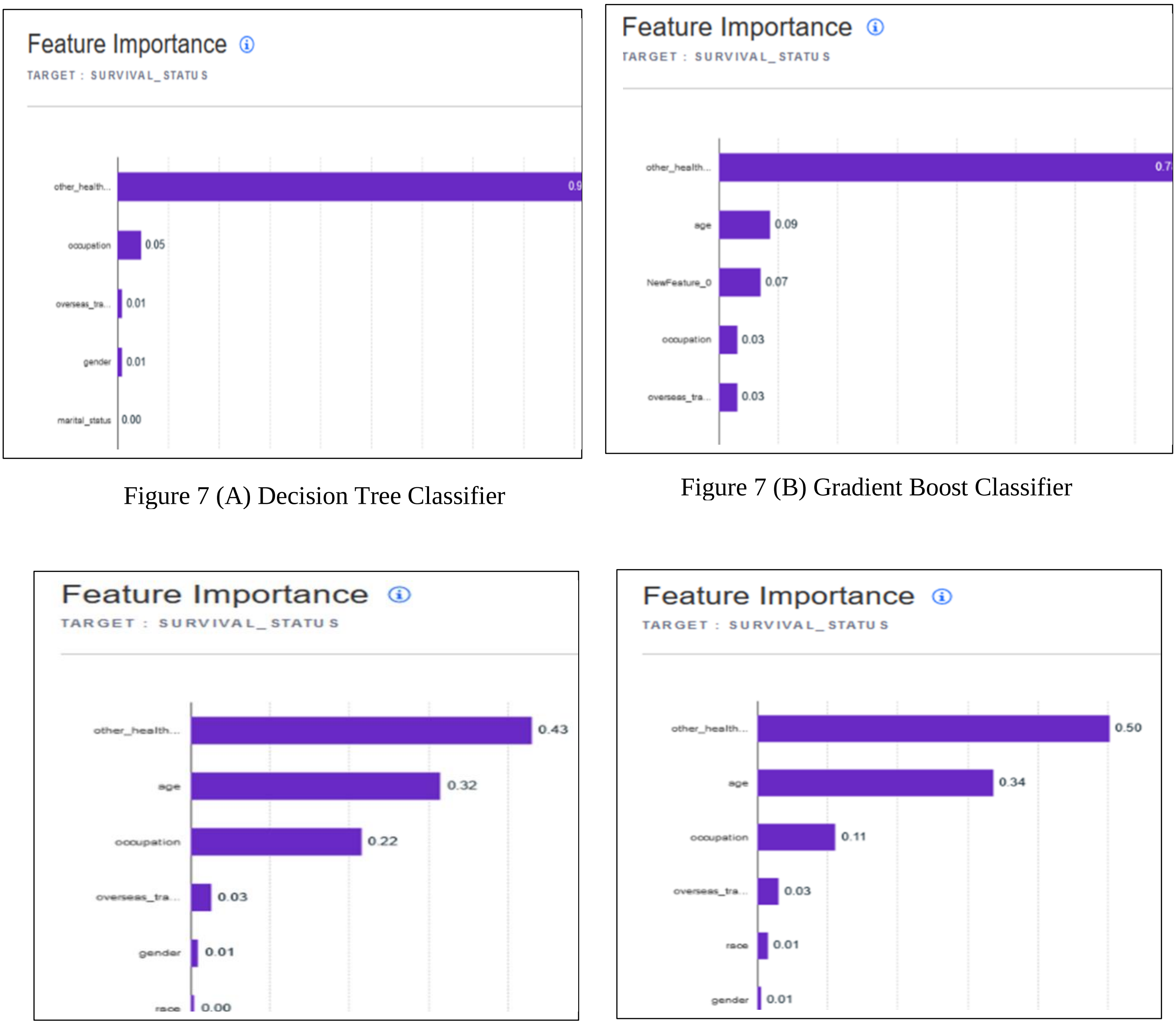

Figure 7 (A) Decision Tree Classifier

Figure 7 (B) Gradient Boost Classifier

Figure 7 (C) Random Forest Classifier

Figure 7 (D) Logistic Regression

Figure 7. Comparison of Feature Performance of COVID-Survivability

4.0 DISCUSSION

This study implemented machine learning models using the COVID-19 dataset as on the 29th April 2020, from the Nigerian Center for Disease Control (NCDC) to identify the most important factors responsible for the survival of infected patients. Of the four chosen machine learning models, three (Decision trees, random forest, and gradient boosting algorithms) yielded prediction accuracies of over 95% with logistic regression with 70% accuracy. The models also revealed the two most important factors that determine patients survivability, and these are underlying health conditions and age of the patients, Patients' occupation and education were distant far from the top two. At the same time, gender, race, travel history, and marital status did not influence patients’ survivability.

Considering the increasing need for predictive medicine and the rising dependence on models of machine learning and data science, this work presents this approach in the study of the current outbreak of the coronavirus that has brought unprecedented difficulties and for which there is still no known cure or vaccines. The intent is to identify the most influencing factors responsible for fatalities among patients (Figure 7) while demonstrating the usability of clinical data as training datasets for different types of machine learning algorithms and comparatively analyzing their efficiencies.

Since the objective of the research was to develop machine learning models that predicted the survivability among COVID-19 patients using clinical data sourced from the Nigerian center for disease control, it is crucial to consider the efficiencies of the chosen algorithms. The performance of each algorithm is evaluated using the receiver operator characteristic (ROC) curve, the F1 score, average precision, and log loss. Table 4. Furthermore, in terms of accuracy during testing with blinded datasets, the reliability of the models showed promising results, the logistic regression model reported the lowest accuracy (78.6), this is followed by Decision tree classifier (95.5), the random forest (96.4). The Gradient boosting algorithm reported 99.3% correct prediction making it the most reliable of all the models. One of the significant strengths of this work, therefore, was the use and comparison of different machine learning classification algorithms to determine the model with the best performance.

The accuracies of the four models on the sample of the dataset are presented in Table 4. The feature importance of all the models is shown in figure 7. The gradient boosting model, random forest, and decision tree all indicated well-calibrated predictions as their curve was almost diagonal; this is not the case with the linear regression model. The COVID-19 clinical dataset appeared to be sufficiently reliable as the calibration measures were close to the identity. The highest accuracy is found with the gradient boosting algorithm (99%). The training dataset, which is 70% of the entire dataset, was used to train and fit the variables. Once the model was processed using the training dataset, predictions were made using the testing dataset (30%). To avoid over-fitting, the validation dataset stopped training as errors increased. As such, the training set indicated an error rate of 0.4 – 0.5, while the testing data indicated an error rate of

0.1-0.3 during prediction. The summary of the models' outcomes (accuracies and performance metrics) are presented in Tables 3 and 4.

5.0 CONCLUSİON

This study has presented a predictive model for the survivability of COVID-19 patients using machine learning, which is a distinction from disease diagnostic systems. Predicting survivability involves efforts towards determining the outcome after an individual has been infected, and this is helpful for a better understanding of the risk factors. In this study, we identified significant predictors of survival of COVID-19 patients using four machine learning models trained with clinical data. This provides evidence-based information, and the system can hence serve as decision support for better understand and individualize hospital management of patients of COVID-19 to improve survival rate.

The research also compares and assesses the performance of four different machine learning algorithms to determine the most efficient algorithm; the gradient boosting machine learning algorithm showed the best results when compared to decision trees, random forest, and logistic regression models. The result reveals that machine learning methods can be effectively utilized in the prediction of survivability in diseases that rely on several factors and promises higher accuracies when compared to conventional statistical or expert-based systems.

Furthermore, the study reveals the two most important variables for patients’ survivability; these are underlying health conditions and age. These findings aligned with the long-held scientific belief that patients with underlying health conditions hardly survived such Pandemic infections. Though the result of these models showed high accuracies in prediction, further studies could consider extending the data set to other continents and get datasets from different countries and different ethnicities. In such cases, environmental conditions and geo-political reasons could be considered to reveal other factors that may be based on the ethnicity or geo-economic analysis for the survivability of COVID-19 patients.

REFERENCES

1. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A, Diaz G. First case of 2019 novel coronavirus in the United States. New England Journal of Medicine. 2020 Jan 31.
2. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, Tan KS, Wang DY, Yan Y. The origin, transmission, and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Military Medical Research. 2020 Dec;7(1):1-0.
3. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery. 2020 Feb 26
4. Burrell CJ, Howard CR, Murphy FA. Coronaviruses. Fenner and White's Medical Virology. 2017:437.
5. Metsky HC, Freije CA, Kosoko-Thoroddsen TS, Sabeti PC, Myhrvold C. CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design. bioRxiv. 2020 Jan 1.
6. World Health Organization. Coronavirus disease 2019 (COVID-19): situation report, 57.
7. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar, L, Favre G, Real estimates of mortality following COVID-19 infection. *Lancet Infect. Dis.* 2020. https://doi.org/10.1016/S1473-3099(20)30195-X
8. Li X, Geng M, Peng Y, Meng L, Lu S. Molecular immune pathogenesis and diagnosis of COVID-19. Journal of Pharmaceutical Analysis. 2020 Mar 5.
9. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Current Biology. 2020 Mar 19.
10. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of autoimmunity. 2020 Feb 26:102433.
11. McKibbin WJ, Fernando R. The global macroeconomic impacts of COVID-19: Seven scenarios.
12. Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A. Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. Available at SSRN 3550308. 2020 Mar 5.
13. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2006 Jan;2:117693510600200030.

14. Soni J, Ansari U, Sharma D, Soni S. Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering. 2011 Jun;3(6):2385-92.
15. Eyck J V, Zadeh M K, Rezapour M, Al-Hyari A Y, Song X, Qiu Z, et al . Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. *Semantic Scholar* 2016.
16. Sriram TV, Rao MV, Narayana GS, Kaladhar DS, Vital TP. Intelligent Parkinson disease prediction using machine learning algorithms. International Journal of Engineering and Innovative Technology (IJEIT). 2013 Sep;3(3):1568-72.
17. Simes RJ. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. Journal of chronic diseases. 1985 Jan 1;38(2):171-86.
18. Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. Journal of medical systems. 1991 Feb 1;15(1):11-9.
19. Cicchetti DV. Neural networks and diagnosis in the clinical laboratory: state of the art. Clinical chemistry. 1992 Jan 1;38(1):9-10.
20. Petricoin EF, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. Current Opinion in Biotechnology. 2004 Feb 1;15(1):24-30.
21. Bocchi L, Coppini G, Nori J, Valli G. Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. Medical Engineering & Physics. 2004 May 1;26(4):303-12.
22. Benza R L, Miller D P, Gomberg-Maitlan M, Frantz R. P, Foreman A J, Coffey C S, et al Predicting Survival in Pulmonary Arterial Hypertension Insights From the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL). *America Heart Association (AHA) Journals* **http://circ.ahajournals.org**, 2010 DOI: 10.1161/circulationaha.109.898122
23. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC medical informatics and decision making. 2019 Dec 1;19(1):48.
24. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005 Jun 1;34(2):113-27.
25. Mihaylov I, Nisheva M, Vassilev D. Application of machine learning models for survival prognosis in breast cancer studies. Information. 2019 Mar;10(3):93.

26. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Computer methods and programs in biomedicine. 2018 Jan 1;153:1-9.

27. Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, Ahmad LG. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013 Apr;4(2):124.

28. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17.

29. Bind S, Tiwari AK, Sahani AK, Koulibaly PM, Nobili F, Pagani M, Sabri O, Borght TV, Laere KV, Tatsch K. A survey of machine learning based approaches for Parkinson disease prediction. International Journal of Computer Science and Information Technologies. 2015;6(2):1648-55.

30. Medhekar DS, Bote MP, Deshmukh SD. Heart disease prediction system using naive Bayes. Int. J. Enhanced Res. Sci. Technol. Eng. 2013 Mar;2(3).

31. WANG L, WANG L. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities.

32. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technology and Health Care. 2016 Jan 1;24(1):31-42.

33. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, Schuchter LM, Shulman LN, Navathe AS, Patel MS, O'Connor NR. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. JAMA network open. 2019 Oct 2;2(10):e1915997-.

34. Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. International journal of medical informatics. 2017 Jan 1;97:304-11.

35. Gupta S, Tran T, Luo W, Phung D, Kennedy RL, Broad A, Campbell D, Kipp D, Singh M, Khasraw M, Matheson L. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ open. 2014 Mar 1;4(3):e004007.

36. Wickham H. and Grolemund G. R for Data Science: *Journal of Statistical Software* ISBN 978-1-4919-1039-9. 522 2017 pp. http://r4ds.had.co.nz/

37. Tanasa D, Trousse B. Advanced data preprocessing for intersites web usage mining. IEEE Intelligent Systems. 2004 Mar;19(2):59-65.

38. Dash M, Liu H. Feature selection for classification. Intelligent data analysis. 1997 Jan 1;1(3):131-56.

39. Mitchell TM. Machine learning.

40. Ijegwa AD, Olufunke VR, Folorunso O, Richard JB. A Bayesian based system for evaluating customer satisfaction in an online store. InProceedings of SAI Intelligent Systems Conference 2018 Sep 6 (pp. 1047-1061). Springer, Cham.

41. Alpaydin E. Introduction to machine learning. MIT press; 2020 Mar 17.

42. Michie D, Spiegelhalter DJ, Taylor CC. Machine learning. Neural and Statistical Classification. 1994 Feb 17;13(1994):1-298.

43. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007 Jun 10;160:3-24.

44. Ayodele TO. Types of machine learning algorithms. New advances in machine learning. 2010 Feb 1:19-48.

45. Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar 1;1(1):81-106.

46. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002 Dec 3;2(3):18-22.