World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Prediction of post-covid-19 using supervised machine learning techniques

Sunday Akinwamide *, Rashidat Idris-Tajudeen and Titilope Helen Akin-Olayemi

Department of Computer Science, The Federal Polytechnic, Ado Ekiti, Nigeria.

World Journal of Advanced Engineering Technology and Sciences, 2024, 12(02), 355–369

Publication history: Received on 02 June 2024; revised on 20 July 2024; accepted on 22 July 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.12.2.0297

Abstract

The COVID-19 pandemic has had a profound impact on global health, necessitating the development of predictive models to manage and mitigate its effects. Early diagnosis is crucial for preventing the progression of diseases that can significantly endanger human life. This study explores the application of supervised machine learning techniques to predict Post-COVID-19 outcomes, including long-term health complications and recovery trajectories. In this study, we utilized 10 advanced supervised machine learning algorithms, including both stand-alone models (Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Gaussian Naive Bayes) and ensemble learning techniques (Bagging Decision Tree Ensemble, Boosting Decision Tree Ensemble, Voting Ensemble, and Stacked Generalization – Stacking Ensemble). These models were applied to analyze and predict the presence of COVID-19 using the COVID-19 Symptoms and Presence dataset from Kaggle. The performance of each model was evaluated using an 80:20 train-test split as well as 5, 10, 15, 20, and 25-fold cross-validation. Evaluation metrics included accuracy, precision, recall, F1-score, and the confusion matrix. The results indicate that the Decision Tree algorithm outperformed the other models, achieving an accuracy of 98.81%, a precision of 1.00, a recall of 0.98, and an F1-score of 0.99. Our results indicate that machine learning models can effectively predict Post-COVID-19 conditions, providing valuable insights for healthcare providers and policymakers.

Keywords: COVID-19; Supervised Machine Learning; Predictive Modeling; Long-term Health Outcomes; Post-COVID Conditions

1. Introduction

The causative agent of the alarming health condition called COVID-19 is an emerging coronavirus that began in Wuhan, China, last December 2019. The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has led to unprecedented challenges in healthcare systems worldwide. The respiratory tract of a human shall be infected, and normally, certain people, especially those with very strong immune systems, can recover over time without special needs for health care (WHO, 2024). But for some, it is going to be completely different. People like the elderly, especially those with diabetes, cancer, respiratory issues, or heart problems that already exist are at an extremely high risk. COVID-19 is a multi-systemic disease and is not actually a respiratory disease. New research has revealed the systemic inflammatory impact of the virus, which was shown to activate nearly every organ in the body (Temgoua et al., 2020). Furthermore, approximately 10–15% of COVID-19 patients develop severe symptoms. These individuals may experience long COVID-19, which can lead to complications affecting the heart, lungs, and nervous system (Ames, 2024). COVID-19 is transmitted through infected airborne droplets during speaking, coughing, and sneezing, or by touching things and being in a contaminated environment. World Health Organization (WHO) insists that not touching one's face, wearing a mask, social distancing, and handwashing can definitely help to avoid being infected. The agency had identified numerous symptoms. The organization justified that from the registered cases, the three most symptomatic were fever, cough, and tiredness, while the six symptoms of the virus were considered to be consequent, namely headache, aching throat, diarrhea, conjunctival congestion, loss of taste or smell, and a skin rash, whilst the symptoms experienced were

* Corresponding author: Sunday Akinwamide; Email: akinwamide_so@fedpolyado.edu.ng

of severe kind with breathing difficulties or shortness of breath, chest pain and loss of speech or movement. There were 182,333,954 cases of COVID-19 with 3,948,238 deaths worldwide as of June 29, 2021 (Worldometer, 2024). The disease has mutated into several variants documented in the US, UK, South Africa, India, and Brazil, among other countries, which carry such differences as increased death rates, quicker transmission, and reduced efficacy of vaccines (CDCP, 2024). A COVID-19 outbreak could lead to an extreme shortage of medical supplies and medical personnel, and of course, COVID-19 testing kits, as the virus continued to spread despite community efforts (Wynants et al., 2020). Inadequate supply of COVID-19 testing kits can hinder early diagnosis of the disease, making it challenging to provide the best possible care for suspected COVID-19 patients. Consequently, there is a need to design the automated prediction system for the identification of whether an individual has COVID-19 or not. Designing a COVID-19 prediction model requires machine learning classification algorithms, datasets, and relevant software. Machine learning can be categorized into supervised, unsupervised, and reinforcement learning. Supervised learning trains machines using labelled datasets, where examples are correctly labelled according to their respective classes (Supervised Versus Unsupervised Learning, 2024). The machine now will predict future occurrences out of the given data, of which the predictions were made based on exactly what was learned from previous data. Unlike supervised machine learning, unsupervised machine learning learns by itself without the presence of the correctly labelled data. The machine will be fed with the data in unsupervised, and the job of the machine is to learn these underlying patterns imbedded within the dataset. In reinforcement learning, the machine functions as an agent that seeks to determine the most appropriate actions through a trial-and-error approach and by observing its environment (Kaelbling & Littman, 1996). Each time the machine successfully performs a task, it receives a reward that increases its state; if it fails, its state is decreased as a form of punishment. This process is repeated multiple times until the machine learns to perform the task correctly. Reinforcement learning is used to train robots to perform human-like tasks and provide personal assistance.

While much focus has been on immediate treatment and containment, understanding and predicting long-term outcomes for recovered patients is crucial. Post-COVID conditions, also known as "long COVID," present a range of symptoms that can persist for months. This paper investigates the use of supervised machine learning techniques to predict these long-term outcomes, aiming to enhance patient care and resource allocation.

The rest of the paper is organized as follows: Section 2 provides an extensive literature review. Section 3 details the framework and thoroughly describes our proposed methods. Section 4 presents the experimental results, and Section 5 offers the discussion and conclusion.

2. Literature Review

Several studies have explored the immediate impacts of COVID-19, but research on long-term outcomes is still emerging. Recent advancements in machine learning (ML) have shown promise in predictive healthcare, particularly in identifying risk factors and outcomes based on patient data. We review existing literature on post-COVID conditions and machine learning applications in healthcare, highlighting gaps that this study aims to fill. This section provides a comprehensive overview of the advancements and contributions made, including an examination of the state-of-the-art methods outlined in prior research concerning COVID-19 detection. In order to gain a deeper understanding of the current study's standing within the realm of Artificial intelligent (AI) contributions, particularly focusing on the utilization of ML algorithms, we conducted a thorough review of pertinent literature regarding COVID-19 detection. This review emphasizes the methods employed, the contributions made, and any associated limitations.

The work of (Villavicencio et al., 2021) proposed an automatic prediction of COVID-19 in a person utilizing J48 DT, RF, SVM, k-NN, and NB algorithms, by analyzing COVID-19 symptoms using Waikato Environment for Knowledge Analysis (WEKA). Each model's performance was evaluated using 10-fold cross validation and compared according to major accuracy measures, correctly or incorrectly classified instances, kappa, mean absolute error, and time taken to build the model. The researchers used the COVID-19 Symptoms and Presence dataset from Kaggle. The results show that SVM using Pearson VII universal kernel outweighs other algorithms by attaining 98.81% accuracy and a mean absolute error of 0.012.

In a related development, (Abayomi-Alli et al., 2022) investigated and applied ensemble learning approach to develop prediction models for effective detection of COVID-19 using routine laboratory blood test results by comparing the performance of different state-of-the-art ML models. The dataset used in this study contains 279 cases of patients from San Raffaele Hospital Milan, Italy. The dataset consists of the results of the respiratory tract RT-PCR test of the samples for 177 positively established cases of COVID-19 and 102 non-COVID-19 cases based on the asopharyngeal swab. Their findings show that an ensemble learning model based on DNN and ExtraTrees achieved a mean accuracy of 99.28%, Area under curve (AUC) of 99.4%, while AdaBoost gave a mean accuracy of 99.28% and AUC of 98.8%. Bashar and his team (Bashar et al., 2021) proposed, implemented, and tested an enhanced augmented normalized chest X-ray image

dataset with the use of optimized DL models, namely, VGG19, VGG16, DenseNet, AlexNet, and GoogleNet. A publicly available dataset of chest X-rays on Kaggle consisting of 21,165 anterior-to-posterior and posterior-to-anterior chest X-ray images classified as: Normal (48%), COVID-19 (17%), Lung Opacity (28%) and Viral Pneumonia (6%) was used in their study. The highest classification accuracy of 95.63% through the application of VGG16 was achieved with non-freeze weights and enhanced normalized augmented data. Mujeeb and his team (Mujeeb et al., 2021) suggested a technique to estimate the likelihood of COVID-19 contamination using a dataset of X-ray images from both healthy individuals and COVID-19 patients. This approach was assessed through various experimental analysis metrics, including accuracy, precision, recall, and F1-score. The experimental outcomes demonstrated that the proposed method could predict COVID-19 presence with an accuracy exceeding 97%. In another research, In 2021, Choudary (Choudary et al., 2021) employed four different machine learning approaches to determine whether COVID-19 was present in each patient. At 98.38%, the SVM algorithm yielded the highest accuracy while Khanday (Khanday et al., 2020) proposed classification of textual clinical reports into four different categories of diseases, namely: COVID-19, SARS, ARDS and both COVID-19 & ARDS using traditional and ensemble machine learning algorithms. The traditional machine learning algorithms employed include: SVM, multinomial NB, LR, and DT while ensemble machine learning techniques used include: RF, bagging, Ada Boost and Stochastic Gradient Boosting. After performing all necessary pre-processing and feature engineering on the dataset, LR and multinomial NB classifiers gave excellent results by having accuracy of 96.2%, 94% precision, 96% recall and 95% f1 score.

The authors, (Muhammad et al., 2021) proposed supervised machine learning predictive models for COVID-19 infection using DT, LR, NB, SVM and ANN. The correlation coefficient analysis between various independent and dependent features was carried out to determine a strength relationship between each independent feature and dependent feature of the dataset prior to developing the models. The result of the performance evaluation of the models showed that DT has the highest accuracy of 94.99% while the SVM has the highest sensitivity of 93.34% and NB has the highest specificity of 94.30%. In another moment, a COVID-19 identification method based on XGBoost, LDA, LR, RF, and Decision Tree machine learning models was proposed by (Zheng et al., 2020). The influence of selecting specific features and variables as well as reducing the dimensionality of features from 12 to 4 were examined by the authors. They came to the conclusion that XGBoost produced the best accuracies for 12-variable and 4-variable models, with 89.6% and 85.9%, respectively.

Rasheed and other researchers (Rasheed et al., 2021) augmented a comparatively bigger X-ray picture dataset to a total of 500 photos, with 408 images overall—of which 50% are COVID-19 positive. CNN and logistic regression were the two classification models that were taken into consideration. 95.2% and 97.6% accuracy rates, respectively, were attained by these models. In a different research, (Bao et al., 2020) examined the identification of COVID-19 on regular blood tests using two machine learning methods. The scientists employed RF and SVM machine learning algorithms on a limited dataset consisting of 294 blood samples sourced from Kunshan People's Hospital and Wuhan Union Hospital in China. SVM beat random forest classifiers with accuracy, precision, sensitivity, and specificity of 84%, 92%, 88%, and 80%, respectively while 15 features were chosen for investigation. (Brinati et al., 2020) examined a different dataset comprising 279 cases from San Raffaele Hospital in Milan, Italy, in order to assess its potential for early COVID-19 identification. The experimental results demonstrated that the RF model beat other classifiers with an accuracy of 86% and a sensitivity of 95% in the performance of seven machine learning models, including KNN, DT, NB, extremely randomized trees (ET), LR, RF, and SVM. The authors of (Ahammed et al., 2020) conducted a more thorough study in which they employed a total of 17 different ML- and DL-based classifier types on a dataset of size 2905 images. This dataset included 219 cases related to COVID-19, 1324 normal cases, and 1362 cases of viral pneumonia. The classifiers applied were CNN, XGB, DNN, ResNet50, VGG16, InceptionV3, SVM, k-NN, GNB, BNB, DT, LR, RT, GB, XGB, NC, and MLP. With an overall accuracy of more than 94%, the CNN model demonstrated the highest level of accuracy performance.

Subsequent research (Barbosa et al., 2022) examined and used six cutting-edge techniques, including Bayesian networks (BN), MLP, SVM, RT, NB, and RF. A dataset of 564 samples, including 559 established COVID-19 samples from Brazil's Albert Einstein Hospital, was used for the investigation. Due to the small amount of data, the authors oversampled using the SMOTE methodology, and they used two algorithms based on PSO and evolutionary search in addition to a manual method for feature selection. The BN model produced the performance model with the best outcomes, with the following metrics: sensitivity, 93.8%, 93.6%, and 95.159% for accuracy, precision, and sensitivity, respectively. According to (Kukar et al., 2021), the authors used an extreme gradient boosting (XGBoost) model to determine the COVID-19 virus. Obtainable from the University Medical Center in Ljubljana, Slovenia, were 5333 blood samples in total, 160 of which were established COVID-19 samples. The experiment's results demonstrated an enhanced AUC of 97%, 81.9% sensitivity, and 97.9% specificity after 35 relevant features were chosen for additional examination. Using the dataset of 598 blood samples from the Albert Einstein Hospital in Brazil, (Banerjee et al., 2020) assessed the blood test results to conduct an initial screening of patients who were likely to have COVID-19. There are 81 COVID-19 instances in the dataset. The authors used machine learning (ML) models based on random forest, logistic regression,

artificial neural network (ANN), and Lasso elastic-net regularized generalized linear network (GLMNET) to base their experiment on 14 blood characteristics. The highest-performing model provided an accuracy of 87% for ANN. A machine learning algorithm that can identify if a person has COVID-19 and is at risk of developing acute respiratory distress syndrome (ARDS) was proposed by (Jiang et al., 2020). 80% accuracy was achieved with the suggested model. Only two Chinese hospitals' worth of patient samples—53 in total—were utilized to train their algorithm.

Nevertheless, some of the shortcomings of the reviewed studies include:

- challenges of limited dataset samples and imbalance datasets
- too much (outlier and noisy) data
- insufficient clinical data that are useful to improve model classification,
- incomplete records, inconsistent data collection methods, and privacy concerns
- bias and variability in training data leading to skewed predictions.

Therefore, identifying some existing feature selection methods for the purpose of dimension reduction is very significant towards these successful classification models (Pulkit, 2024). Additionally, research should focus on analyzing the integrated performance of new test data using various ML algorithms (Khekare et al., 2022). Addressing some of the limitations of previous studies, this research applied and assessed the performance of various advanced machine learning algorithms, encompassing both standalone models and ensemble learning techniques. The dataset was divided using an 80:20 train-test split and also subjected to 5, 10, 15, 20, and 25-fold cross-validation to ensure effective COVID-19 prediction.

3. Proposed methodology

The machine learning modelling process was conducted in the Anaconda Distribution Integrated Development Environment, utilizing Python in a Jupyter Notebook. This section provides a thorough description of the proposed experimental model, with a visual overview shown in Figure 1. Our study applied and evaluated the performance of various state-of-the-art machine learning algorithms, including Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Gaussian Naive Bayes, Bagging Decision Tree Ensemble Learning, Boosting Decision Tree Ensemble Learning, Voting Ensemble Learning and Stacked Generalization – Stacking Ensemble Learning. These were all assessed using both an 80:20 train-test split and 5, 10, 15, 20, and 25-fold cross-validation for effective COVID-19 prediction.

3.1. Dataset Collection and Description

The dataset comprises clinical and demographic information from patients who have recovered from COVID-19. Data sources include hospital records, follow-up surveys, and public health databases. The dataset utilized for this research is titled “Symptoms and COVID Presence”, accessible from Kaggle at (Hemanth, 2024). It comprises 20 features representing potential factors associated with contracting the virus and 1 target variable indicating the presence of COVID-19. The dataset comprises 5,435 instances and 21 attributes. Out of these, 4,383 instances are labelled as "Yes", indicating the presence of COVID-19, while the remaining 1,051 instances are labelled as "No", indicating the absence of COVID-19. All records in the dataset are composed of categorical variables, and there are no missing values.

3.2. Dataset Pre-processing

Data pre-processing is a crucial step in machine learning because the quality of data and the information extracted from it directly impact the model's learning capability. Therefore, it is essential to pre-process our data before feeding it into the model (Dhairya, 2024). Data pre-processing steps involve handling missing values, normalizing features, and encoding categorical variables. We also applied techniques such as SMOTE to address class imbalance in the dataset.

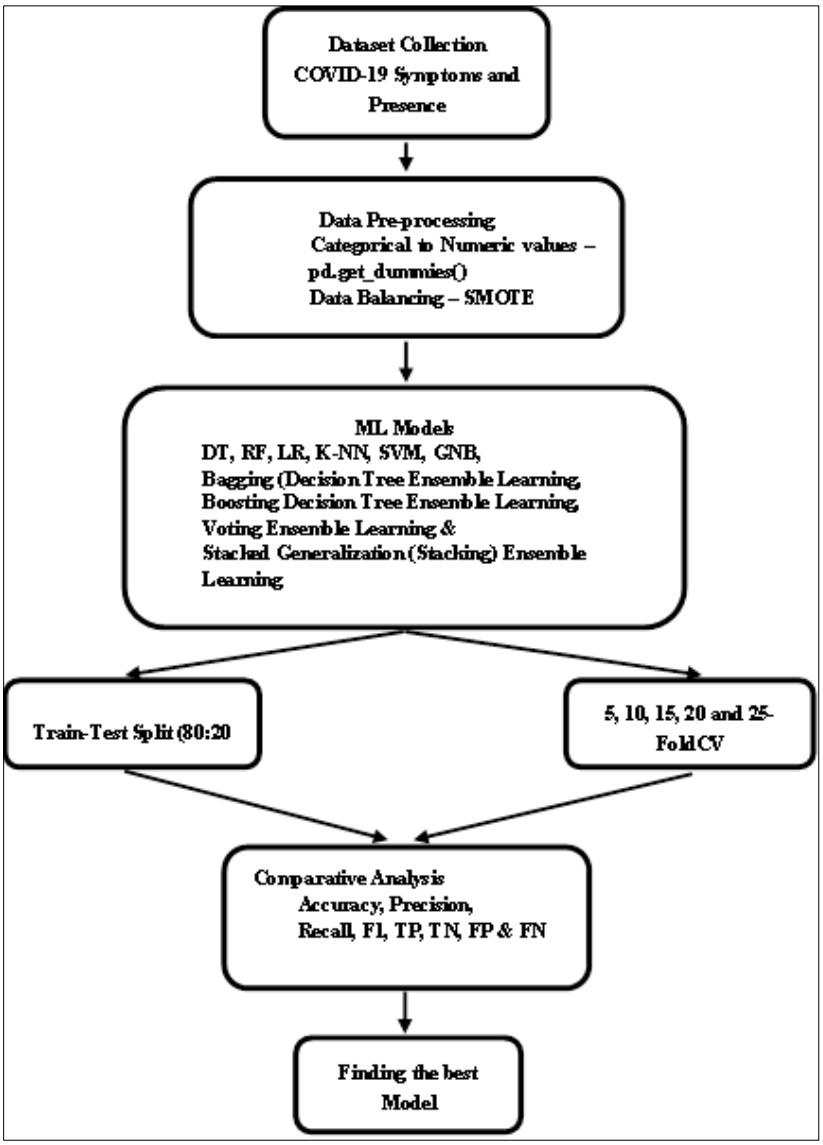

Figure 1 Graphical Overview of the Proposed Methodology

3.2.1. Converting Categorical Data to Numerical Equivalent

Initially, we converted the categorical target variable "COVID-19" into numerical values, by assigning 0 to the "No" class and 1 to the "Yes" class. The remaining independent variables were transformed into a numerical format using the pd.get_dummies function from the panda's library. This function converts categorical variables into a format that machine learning algorithms can process effectively.

3.2.2. Feature Selection

The feature selection for the independent variables in this research was performed using the feature_importances_ attribute of the Random Forest Classifier object from scikit-learn. Feature importance in a Random Forest can be determined using a metric called Gini importance. This metric measures the total reduction in Gini impurity of the dataset when a particular feature is used for splitting. A higher Gini importance indicates that the feature is more significant for the model (Prasanna, 2024). Figure 2 shows the result of the feature selection.

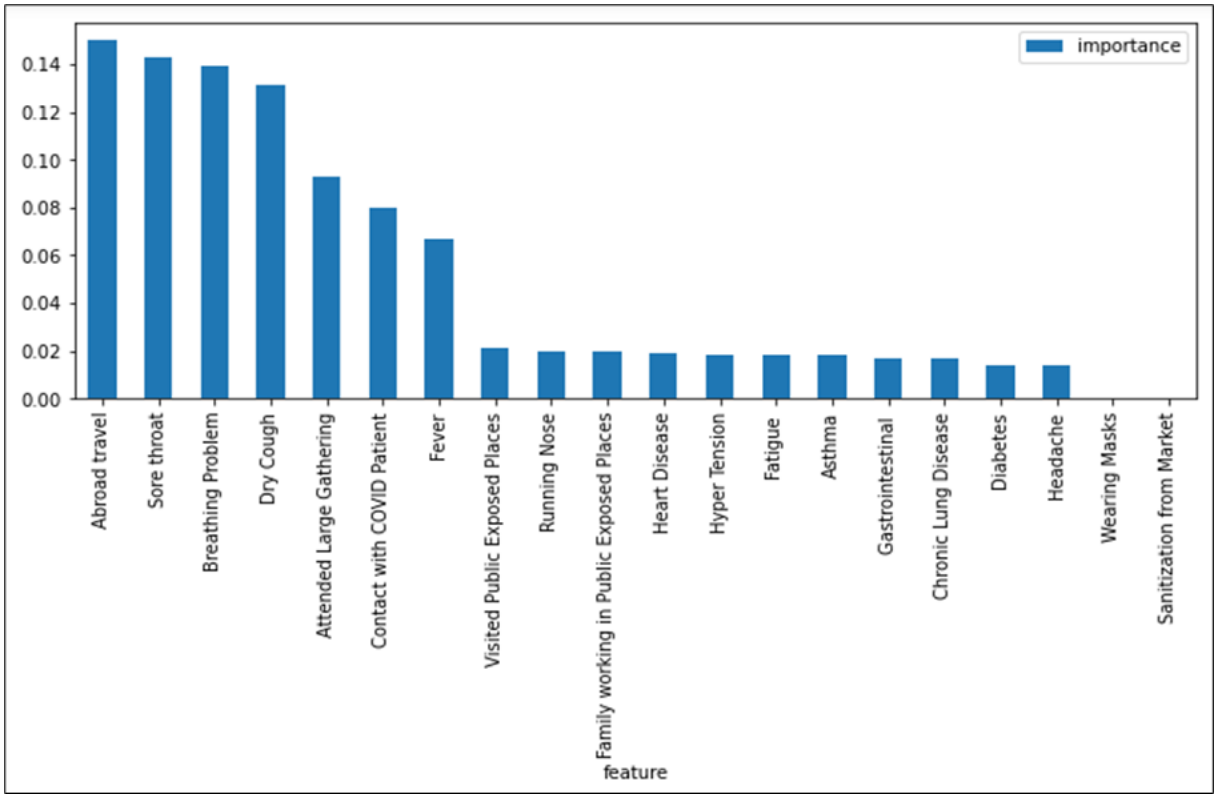

Figure 2 Result of the Feature Selection using the feature_importances_ attribute of the Random Forest Classifier

According to Figure 2, it is evident that the “Wearing Masks” and “Sanitization from Market” attributes do not contribute to the target variable. Therefore, they were dropped, having only 18 features left to train our machine learning models.

3.2.3. Handling Data Imbalance

The issue of data imbalance was identified in our dataset due to the disproportionate ratio between the COVID-19 "Yes" class (4,383 samples) and the "No" class (1,051 instances), approximately a 4:1 ratio. Data imbalance can cause prediction bias, as the large number of samples in the COVID-19 "Yes" class may lead machine learning classifiers to become overly familiar with this class. As a result, there's a significant risk that the model may predominantly predict COVID-19 as "Yes", potentially skewing its predictions (Swastik, 2024). To address this, we utilized a technique known as SMOTE (Synthetic Minority Over-Sampling Technique). SMOTE is an oversampling method that generates synthetic samples for the minority class by interpolating between existing examples from that class (LearnArtificialIntelligence, 2024). This process effectively increases the representation of the class with fewer samples in the dataset. Balancing the dataset in this manner is crucial for achieving a high accuracy rate, minimizing the error rate, and preventing classification bias. After applying SMOTE, the total number of samples in the dataset increased to 8,766. Of these, 4,383 samples belong to the COVID-19 "Yes" class, while the remaining 4,383 samples now belong to the COVID-19 "No" class. With the dataset balanced, we proceeded to splitting it using ratio 8:2, indicating that 80% of the samples were allocated as the training dataset for developing the COVID-19 prediction model, while the remaining 20% were set aside for testing the model's performance. The train_test_split function in Sklearn's model selection module divides data arrays into two subsets: one for training and one for testing. It automates the dataset division process, eliminating the need for manual splitting. By default, Sklearn's train_test_split randomly partitions the data into these subsets. However, you can also set a specific random state for the operation if desired (Train_test_split, 2024). By applying the train_test_split function, 7,012 samples were included in the training dataset, and the remaining 1,754 samples were taken to be used as the testing dataset.

3.3. K-Fold Cross Validation

Cross-validation is a machine learning technique employed to assess a model's performance on unseen data. It entails splitting the available data into numerous folds or subsets, designating one fold as the validation set, and training the model on the remaining folds. This cycle is iterated multiple times, each time utilizing a different fold as the validation set. Ultimately, the outcomes from each validation iteration are averaged to generate a more reliable estimate of the model’s performance. Cross-validation is a crucial phase in the machine learning workflow, ensuring that the chosen

model for deployment is robust and capable of generalizing effectively to new data (Geeks for Geeks, 2024). In this study, aside from training and testing the 10 selected ML algorithms using an 80:20 split with the train_test_split function from Sklearn, we also evaluated the performance of our proposed model through 5-fold, 10-fold, 15-fold, 20-fold, and 25-fold cross-validation tests across all the chosen ML models using the cross_val_predict() function from sklearn. This step was taken to provide a dependable evaluation of the model's performance and prevent overfitting. All the outcomes were then compared to identify the best-performing model to predict COVID-19.

3.4. Modelling

Our research delves into evaluating the efficacy of 10 cutting-edge ML algorithms, encompassing stand-alone classifiers and ensemble learning approaches. The different models used in this study are described as follows:

3.4.1. Decision Tree

Decision Trees serve as the basis for several classical machine learning algorithms such as Random Forests, Bagging, and Boosted Decision Trees. The concept was originally introduced by Leo Breiman, a statistician at the University of California, Berkeley. Breiman's concept involved structuring data as a tree, with each internal node representing a test on an attribute (essentially a condition), each branch indicating an outcome of the test, and each leaf node (terminal node) containing a class label (Vihar, 2024).

3.4.2. Support Vector Machine

Support Vector Machines (SVMs) are a supervised machine learning algorithm utilized for both classification and regression tasks. They find extensive application across fields such as pattern recognition, image analysis, and natural language processing. SVMs operate by identifying the optimal hyperplane that segregates data points into distinct classes. This model facilitates predictions on new data points by determining their position relative to the hyperplane. Points on one side are classified as belonging to one class, while those on the opposite side belong to another class (Tasmay, 2024).

3.4.3. Random Forest

Random Forest is like a group a decision-making team in machine learning. It combines the opinions of many “trees” (individual models) to make better predictions, creating a more robust and accurate overall model. Random Forest is a widely-used machine learning algorithm which combines the output of multiple decision trees to reach a single result. The algorithm’s strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks (Prasanna, 2024).

3.4.4. Gaussian Naive Bayes

Naive Bayes represents a probabilistic approach to machine learning, grounded in Bayes’ theorem. Its main application lies in classification tasks like spam detection, sentiment analysis, and document sorting. The "naive" aspect comes from assuming that the features describing instances are conditionally independent, given the class label. This simplification enables efficient and straightforward computation (Nandini, 2024). The conditional probability is defined in the following manner:

$$P(A|B) = P(B|A) \cdot P(A) \,/\, P(B) \tag{1}$$

In the context of Naive Bayes:
$P(A|B)$ is the probability of class A given the features B
$P(B|A)$ is the probability of observing features B given class A
$P(A)$ is the prior probability of class A
$P(B)$ is the probability of observing features B

3.4.5. Logistic Regression

Logistic regression serves as a supervised machine learning technique designed for binary classification tasks by estimating the probability of an outcome or event. It produces a binary outcome, typically expressed as yes/no, 0/1, or true/false, encompassing two potential results. This method examines the connection between independent variables and categorizes data into distinct classes. Logistic regression finds broad application in predictive modelling, determining the likelihood that an instance falls into a particular category. Logistic regression employs a sigmoid

function, also known as a logistic function, to map predictions and their corresponding probabilities. This function is characterized by an S-shaped curve that transforms real values into a bounded range from 0 to 1 (Vijay, 2024). The sigmoid function serves as an activation function in logistic regression and is formally defined as:

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

where,
e = base of natural logarithms
value = numerical value one wishes to transform

3.4.6. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a widely used technique in machine learning for tasks involving classification and regression. It operates on the principle that similar data points tend to share similar labels or values. During training, KNN stores the entire training dataset as a reference. When making predictions, it computes the distance between the input data point and all training examples using a chosen distance metric, such as Euclidean distance. The algorithm then identifies the K nearest neighbors to the input data point based on these distances. For classification tasks, KNN assigns the most common class label among the K neighbors as the predicted label. For regression tasks, it predicts the value for the input data point by calculating the average or weighted average of the target values of the K neighbors (Tavish, 2024a).

3.4.7. Bootstrap Aggregation (Bagging) Decision Tree Ensemble Learning

Techniques like Decision Trees may be susceptible to overfitting on the training data, resulting in inaccurate predictions on new data. Bootstrap Aggregation (bagging) is an ensemble method designed to address overfitting in classification or regression tasks. Its goal is to enhance the accuracy and performance of machine learning models. Bagging achieves this by creating random subsets of the original dataset, with replacement, and then training a classifier (for classification) or regressor (for regression) on each subset. The predictions from these subsets are combined through majority voting for classification or averaging for regression, thereby improving prediction accuracy (Machine Learning, 2024).

3.4.8. Boosting Decision Tree Ensemble Learning

Boosting is another robust ensemble method that constructs a series of decision trees sequentially. In contrast to bagging, boosting aims to minimize both bias and variance by iteratively training weak learners and assigning greater importance to incorrectly classified instances. Each subsequent tree in the ensemble is trained to rectify the errors made by the previous trees. Boosting algorithms like AdaBoost, Gradient Boosting, and XGBoost have shown impressive performance across different domains, making them favored options for enhancing decision tree accuracy (Sankhyana, 2024).

3.4.9. Voting Ensemble Learning

Voting stands as a core component of ensemble learning, leveraging predictions from multiple models to reach a consolidated outcome. Within ensemble techniques, two primary voting methods emerge: hard voting and soft voting. Hard voting, also termed majority voting, centers on aggregating predictions from base models and selecting the class with the highest number of votes as the ultimate prediction. This approach suits classification tasks, especially when dealing with discrete and mutually exclusive classes. Soft voting, also referred to as weighted voting, incorporates probability scores from each base model across all classes. It computes the weighted average of these probabilities to formulate the final prediction. In soft voting, the winner is determined by identifying the class with the highest weighted probability, making it adaptable for both classification and regression scenarios (Awan, 2024).

3.4.10. Stacked Generalization (Stacking) Ensemble Learning

Stacking, alternatively known as Stacked Generalization, is an ensemble methodology that merges multiple classification or regression models through a meta-classifier or a meta-regressor. The foundational models are trained on a comprehensive training set, after which the meta-model is trained on the features generated by the base-level model. Stacking ensembles often exhibit heterogeneity due to the diverse learning algorithms employed in the base-level. The base models in stacking, also called Base-Models, are typically varied (e.g., not limited to decision trees) and are trained on the same dataset. Conversely, a single model known as the Meta-model is utilized to optimize the combination of predictions from the contributing models. The architecture of a stacking model comprises two or more base models, commonly referred to as level-0 models, and a meta-model, also called a level-1 model, which aggregates

the predictions of the base models. The meta-model learns to combine the predictions made by the base models on out-of-sample data, with the outputs from the base models serving as input for the meta-model. These outputs may include real values for regression tasks, and probability values, probability-like values, or class labels for classification tasks (Great Learning Team, 2024).

3.5. Evaluation Metrics

Evaluation metrics provide numerical yardsticks applied in the measurement of the performance and effectiveness of a statistical or machine learning model. They give an idea of the model's performance and help in comparing different models or algorithms. Very importantly, machine learning models need to be evaluated from the point of view of predictive accuracy, generalization, and overall quality. They give measurements of such aspects objectively. The choice of evaluation metrics depends on the problem domain, the type of data, and the expected outcomes (Tavish, 2024b). The performance evaluation metrics chosen for this research are as follows:

3.5.1. Confusion Matrix

The confusion matrix is a basic part of a machine learning classification model scoring, though it is not one of the metrics. Theoretically, it is a two-dimensional array that holds actual and predicted values. For example, let's say we have to build a classifier that gives a diagnosis for patients: sick or healthy (AltexSoft, 2024).

Table 1 Confusion Matrix Table

	PREDICTED LABEL		
TRUTH LABEL		0	1
	0	TN	FP
	1	FN	TP

Additionally, each dimension includes instances of such classes as:

- True Positive (TP) — a class is predicted true and is true in reality (patients that are sick and diagnosed sick);
- True Negative (TN) — a class is predicted false and is false in reality (patients that are healthy and diagnosed healthy);
- False Positive (FP) — a class is predicted true but is false in reality (patients that are healthy but diagnosed sick); and
- False Negative (FN) — a class is predicted false but is true in reality (patients that are sick but diagnosed healthy).

3.5.2. Accuracy

Accuracy quantifies the ratio of correct predictions to the total number of predictions, calculated by dividing the number of correct predictions by the total predictions made.

$$\text{Accuracy} = \frac{\textit{Number of Correct Predictions}}{\textit{Total Number of Predictions}} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

3.5.3. Precision

Precision measures the accuracy of positive predictions by determining the proportion of correct positive results (True Positives, TP) divided by the total number of positive results (TP + False Positives, FP) predicted by the classifier.

$$\text{Precision} = \frac{\textit{Number of Correct Positive Results}}{\textit{Total Number of Positive Results}} = \frac{TP}{TP+FP} \quad (4)$$

3.5.4. Recall

Recall indicates the proportion of correctly predicted positive instances out of all the positives that the model could have identified. It's calculated by dividing True Positives (TP) by the sum of True Positives and False Negatives (FN) in the dataset. Unlike precision, which focuses on accurately predicted positives among all positive predictions, recall sheds light on missed positive predictions.

$$\text{Recall} = \frac{\textit{Number of Correct Positives}}{\textit{Number of all Positives}} = \frac{TP}{TP+FN} \quad (5)$$

3.5.5. F1 Score

The F1 Score aims to strike a balance between precision and recall by computing their harmonic mean. It serves as a measure of a test's accuracy, with a maximum value of 1 indicating perfect precision and recall alignment.

$$\textit{F1 Score} = \text{Harmonic Mean of Precision and Recall} = 2 \times \frac{\textit{Precision X Recall}}{\textit{Precision + Recall}} = \frac{2TP}{2TP+FP+FN} \quad (6)$$

3.6. Finding the Best Model

Choosing the best model isn't solely about accuracy; other performance metrics must be considered as well (Brownlee, 2024). Accuracy is most effective when dealing with symmetric datasets or those with similar sample counts per class (Ghoneim, 2024). However, the "Symptoms and COVID Presence" dataset initially had an imbalance, with more COVID-19-positive samples than negatives. Consequently, we evaluated various criteria to select the optimal algorithm for building a COVID-19 predictor. The criteria for identifying the most suitable machine learning algorithm for this predictor are: highest accuracy, precision, recall and F1-score. In addition, we also considered TP, TN, FP and FN of confusion matrix.

4. Results

This section details our findings on the performance of each model, including the experimental values of our evaluation metrics. In addition to training and testing the 10 selected ML algorithms using an 80:20 split, we also assessed the performance of our proposed model through 5-fold, 10-fold, 15-fold, 20-fold, and 25-fold cross-validation tests on all the chosen ML models. The best results of each model are shown in Table 2.

4.1. Comparative Analysis of Results

Table 2 showcases the top performance metrics of the 10 machine learning models evaluated in this study. Although the performance metrics of the first eight models—DT, RF, Bagging, Boosting, Stacking, Voting, K-NN, and GNB—are closely matched, DT clearly stands out with the best overall performance. It achieves an accuracy of 98.81%, a precision of 1.00, a recall of 0.98, and an F1-score of 0.99. A careful examination of Table 2 also reveals that boosting, stacking, and voting exhibit identical performance metrics, each with 98.69% accuracy, a precision of 1.00, a recall of 0.97, and an F1-score of 0.99. Similarly, K-NN and GNB share the same performance metrics, each with 98.52% accuracy, a precision of 1.00, a recall of 0.97, and an F1-score of 0.98, while LR has the lowest performance. Figure 3 provides a visual representation of the models' accuracy results.

Table 2 Best Results of each Model

S/N	Model	Criteria	Acc	Pre	Rec	F1	TP	TN	FP	FN
1.	DT	25-FoldCV	98.81	1.00	0.98	0.99	4279	4383	0	104
2.	RF	20-FoldCV	98.77	1.00	0.98	0.99	4275	4383	0	108
3.	Bagging	20-FoldCV	98.75	1.00	0.98	0.99	4275	4381	2	108
4.	Boosting	80:20	98.69	1.00	0.97	0.99	854	877	0	23
5.	Stacking	80:20	98.69	1.00	0.97	0.99	854	877	0	23
6.	Voting	80:20	98.69	1.00	0.97	0.99	854	877	0	23
7.	KNN	80:20	98.52	1.00	0.97	0.98	851	877	0	26
8.	GNB	5-FoldCV	98.52	1.00	0.97	0.98	851	877	0	26
9.	SVM	80:20	95.95	0.96	0.96	0.96	840	843	34	37
10.	LR	80:20	91.96	0.93	0.91	0.92	797	816	61	80

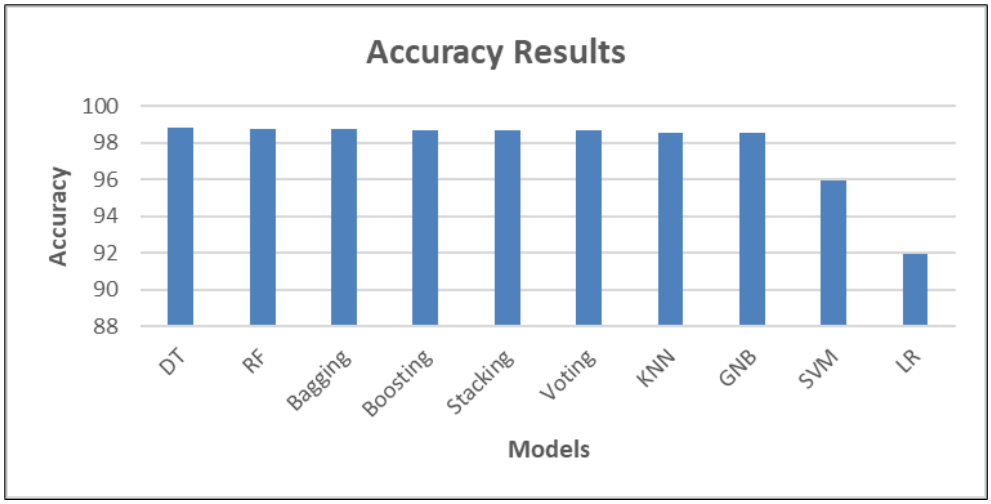

Figure 3 Pictorial Representation of Accuracy Results of the Research

4.2. Comparison of the Proposed Research with Previous Studies

To further evaluate our proposed research, we compared our results with those of previous related studies using the same performance metrics. The proposed study shows a significant improvement compared to the existing studies. Table 3 provides a comparison of related studies and the proposed research on COVID-19 prediction, detailing the datasets used, the most effective algorithms, and the performance evaluation metrics along with their values.

Table 3 Summary of the comparison between related work and the proposed study

Reference	Dataset	Best Algorithm	Metrics
Villavicencio et al., (2021)	COVID-19 Symptoms and Presence from Kaggle	SVM	Acc. 98.81%, Pre. 0.98, Rec. 0.98, F1. 0.98
Abayomi-Alli et al., (2022)	Routine Blood Test from San Raffaele Hospital	Ada Boost and Extra Trees	Acc. 99.28%, AUC. 0.99 & Acc. 99.28, AUC 0.99
Bashar et al., (2021)	Chest X-rays from Kaggle	VGG16	Acc. 95.63%, Pre. 99.18, Rec. 98.78, F1. 98.98
Mujeeb et al., (2021)	X-ray images of both healthy and patients infected with COVID-19	Three-way Random Forest (TWRF)	Acc. 97.6%, Pre. 1.00, Rec. 0.95, F1. 0.97
Choudary et al., (2021)	Symptoms and COVID Presence (May 2020)	SVM	Acc. 98.30%, Rec. 0.99
Khanday et al., (2020)	Clinical Text Data	LR & Multinomial Naïve Bayes	Acc. 96.20%, Pre. 0.94, Rec. 0.96, F1. 0.95
Muhammad et al., (2021)	Epidemiology COVID-19 dataset from Mexico	DT	Acc. 94.99, Rec. 0.89, TNR. 0.93
Bao et al., (2020)	X-rays images of both healthy and patients infected with COVID-19	SVM	Acc. 84.00%, Pre. 0.92, Rec. 0.88, TNR. 0.80
Ahammed et al., (2020)	Chest X-rays images from Kaggle	CNN	Acc. 94.03%, Rec. 0.94, F1. 0.94, TNR. 0.97

Proposed Study	Symptoms and COVID Presence from Kaggle	DT	Acc. 98.81%, Pre. 1.00, Rec. 0.98, F1. 0.99

5. Discussion

The necessity for early and efficient COVID-19 detection methods is crucial during the global pandemic, and utilizing artificial intelligence techniques can greatly enhance prediction accuracy and support health workers in their decision-making process. This research demonstrates the feasibility and clinical reliability of using a dataset containing various symptoms and their outcomes (i.e., whether a person has COVID-19 or not) along with machine learning as alternatives to the commonly used Reverse Transcription Polymerase Chain Reaction (RT-PCR) test for classifying COVID-19 positive patients. This approach is particularly beneficial for countries experiencing shortages of RT-PCR reagents and specialized laboratories, such as those in the developing world.

6. Conclusion

This paper outlines straightforward and intriguing steps for detecting COVID-19 using the Symptoms and COVID Presence dataset from Kaggle. Despite the dataset having no missing values, the study addresses the challenges of extracting significant features that greatly influence the target variable for training the models, as well as dealing with class imbalance.

In our research, besides training and testing the 10 selected machine learning algorithms using an 80:20 dataset split, we evaluated the models through 5-fold, 10-fold, 15-fold, 20-fold, and 25-fold cross-validation tests. The performance accuracy of the top five models, in descending order, are as follows: DT 98.81%, RF 98.77%, Bagging 98.75%, (Boosting, Stacking, & Voting 98.6), and (K-NN & GNB 98.52) respectively. Based on our research findings, we can assert that our proposed methods surpass the current state-of-the-art techniques, as evidenced in Table 3. The ML system for early COVID-19 detection presents a swift, uncomplicated, and more cost-effective substitute to the RT-PCR test.

Our study demonstrates the potential of supervised machine learning techniques in predicting Post-COVID-19 conditions. By leveraging clinical and demographic data, these models can provide valuable insights for healthcare providers, aiding in the management of long-term health impacts. Future research should focus on expanding datasets and refining models to account for emerging variants and evolving patient demographics.

Compliance with ethical standards

Acknowledgments

The authors wish to thank TetFund Nigeria for sponsoring this research work. The Directorate, Centre for Research, Innovation and Development of the Federal Polytechnic, Ado-Ekiti is highly appreciated for their cooperation and understanding. Many thanks to Mrs. Akinwamide Busayo Rachael for her useful suggestions.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Abayomi-Alli, O. O., Damaševičius, R., Maskeliūnas, R., & Misra, S. (2022). An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples. Sensors, 22(6), 2224. https://doi.org/10.3390/s22062224

[2] Ahammed, K., Satu, M. S., Abedin, M. Z., Rahaman, M. A., & Islam, S. M. S. (2020). Early detection of coronavirus cases using chest X-ray images employing machine learning and deep learning approach. 10, 07.

[3] AltexSoft. (2024). AltexSoft 2024. Machine Learning Metrics: How to Measure the Performance of a Machine Learning Model. Available Online: https://www.altexsoft.com/blog/machine-learning-metrics/

[4] Ames, H. (2024). How Long Does Coronavirus Last in the Body, Air, and in Food? Available online: https://www.medicalnewstoday.com/articles/how-long-does-coronavirus-last

[5] Awan, U. R. (2024). Understanding Soft Voting and Hard Voting: A Comparative Analysis of Ensemble Learning Methods. Available Online: https://medium.com/@awanurrahman.cse/understanding-soft-voting-and-hard-voting-a-comparative-analysis-of-ensemble-learning-methods-db0663d2c008

[6] Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., & Mackenzie, L. S. (2020). Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. International Immunopharmacology, 86(106705). https://doi.org/10.1016/j.intimp.2020.106705

[7] Bao, F. S., He, Y., Liu, J., Chen, Y., Li, Q., Zhang, C. R., & Ouyang, L. (2020). Triaging moderate COVID-19 and other viral pneumonias from routine blood tests.

[8] Barbosa, V. A. D. F., Gomes, J. C., De Santana, M. A., Albuquerque, J. E. D. A., De Souza, R. G., De Souza, R. E., & Dos Santos, W. P. (2022). Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests. Research on Biomedical Engineering, 38(1), 99–116. https://doi.org/10.1007/s42600-020-00112-5

[9] Bashar, A., Latif, G., Ben Brahim, G., Mohammad, N., & Alghazo, J. (2021). COVID-19 Pneumonia Detection Using Optimized Deep Learning Techniques. Diagnostics, 11(11), 1972. https://doi.org/10.3390/diagnostics11111972

[10] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. Journal of Medical Systems, 44(8), 135. https://doi.org/10.1007/s10916-020-01597-4

[11] Brownlee, J. (2024). Classification Accuracy is not Enough: More Performance Measures You Can Use, Machine Learning Mastery, 20 June 2019. Available online: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performancemeasures-you-can-use/

[12] CDCP. (2024). Centers for Disease Control and Prevention (CDC). SARS-Cov-2 Variant Classifications and Definitions, 17 May 2021. Available online: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html

[13] Choudary, M. N. S., Bommineni, V. B., Tarun, G., Reddy, G. P., & Gopakumar, G. (2021). Predicting Covid-19 Positive Cases and Analysis on the Relevance of Features using SHAP (SHapley Additive exPlanation). 2021 Second International Conference on Electronics and Sustainable Communication Systems, 1892–1896. https://doi.org/doi: 10.1109/ICESC51422.2021.9532829.

[14] Dhairya, K. (2024). Introduction to Data Pre-processing in Machine Learning. Beginners Guide for Data Pre-processing. Available Online: https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

[15] Geeks for Geeks. (2024). GeeksforGeeks, Sanchhaya Education Private Limited. Cross Validation in Machine Learning. Available Online: https://www.geeksforgeeks.org/cross-validation-machine-learning/

[16] Ghoneim, S. (2024). Accuracy, Recall, Precision, F-Score & Specificity, which to Optimize on? 2 April 2019. Available online: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

[17] Great Learning Team. (2024). Great Learning Team. Ensemble learning with Stacking and Blending. Available Online: https://www.mygreatlearning.com/blog/ensemble-learning/

[18] Hemanth, H. (2024). Symptoms and COVID Presence (May 2020 data). Updated 4 years ago. Available at: https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence

[19] Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., ... &, & Huang, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. 63(1), 537–551. https://doi.org/doi:10.32604/cmc.2020.010691

[20] Kaelbling, L. P., & Littman, M. L. (1996). Reinforcement Learning: A Survey. 4, 237–285.

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

[22] Khekare, G., Turukmane, A. V., Dhule, C., Sharma, P., & Kumar, B., L. (2022). Experimental Performance Analysis of Machine Learning Algorithms. Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications. WCNA 2021. Lecture Notes in Electrical Engineering. https://doi.org/10.1007/978-981-19-2456-9_104

[23] Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., Zalaznik, M., Notar, M., Moškon, S., & Notar, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. Scientific Reports, 11(1), 10738. https://doi.org/10.1038/s41598-021-90265-9

[24] LearnArtificialIntelligence. (2024). LearnArtificialIntelligence. SMOTE: A Powerful Technique for Handling Imbalanced Data. Available Online: https://medium.com/@thecontentfarmblog/smote-a-powerful-technique-for-handling-imbalanced-data-2375ad46103c

[25] Machine Learning. (2024). Machine Learning—Bootstrap Aggregation (Bagging). Available Online: https://www.w3schools.com/python/python_ml_bagging.asp

[26] Muhammad, L. J., Algehyne, E. A., Usman, S. S., Usman, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. 2(11). https://doi.org/10.1007/s42979-020-00394-7

[27] Mujeeb, U. R., Arslan, S., Sohail, K., Maha, D., & Saeed, R. (2021). Future Forecasting of COVID-19: A Supervised Learning Approach. 21(3322). https:// doi.org/10.3390/s21103322

[28] Nandini, V. (2024). Understanding and Implementing Gaussian Naive Bayes Classification with Python. Available Online: https://medium.com/@nandiniverma78988/understanding-and-implementing-gaussian-naive-bayes-classification-with-python-dbdfcf2939f7

[29] Prasanna, G. (2024). Using Random Forest for Feature Importance and Feature Selection. Available Online: https://medium.com/@prasannarghattikar/using-random-forest-for-feature-mportance-118462c40189

[30] Pulkit, S. (2024). The Ultimate Guide to 12 Dimensionality Reduction Techniques (with Python codes). Available Online: https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/

[31] Rasheed, J., Hameed, A. A., Djeddi, C., Jamil, A., & Al-Turjman, F. (2021). A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdisciplinary Sciences: Computational Life Sciences, 13(1), 103–117. https://doi.org/10.1007/s12539-020-00403-6

[32] Sankhyana. (2024). Sankhyana Consultancy Services Pvt. Ltd. Data Driven Decision Science. Available Online: https://www.linkedin.com/pulse/ensemble-techniques-decision-tree

[33] Supervised Versus Unsupervised Learning. (2024). Supervised Versus Unsupervised Learning: Key Differences. Available online: https://www.guru99.com/supervised-vsunsupervised-learning.html

[34] Swastik, S. (2024). SMOTE for Imbalanced Classification with Python. Available Online: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[35] Tasmay, P. T. (2024). Support Vector Machines (SVM): An Intuitive Explanation. Available Online: https://medium.com/low-code-for-advanced-data-science/support-vector-machines-svm-an-intuitive-explanation-b084d6238106

[36] Tavish, S. (2024a). A Complete Guide to K-Nearest Neighbors (Updated 2024). Available Online: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[37] Tavish, S. (2024b). 12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023). Available Online: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

[38] Temgoua, M. N., Endomba, F. T., Nkeck, J. R., Kenfack, G. U., Tochie, J. N., & Essouma, M. (2020). Coronavirus Disease 2019 (COVID-19) as a Multi-Systemic Disease and its Impact in Low- and Middle-Income Countries (LMICs). SN Comprehensive Clinical Medicine, 2(9), 1377–1387. https://doi.org/10.1007/s42399-020-00417-7

[39] Train_test_split. (2024). Train_test_split. Splitting Datasets with the Sklearn train_test_split Function. Available Online: https://www.bitdegree.org/learn/train-test-split

[40] Vihar, K. (2024). An Introduction to Decision Trees. Available Online: https://blog.paperspace.com/decision-trees/

[41] Vijay, K. (2024). What is Logistic Regression? Equation, Assumptions, Types, and Best Practices. Available Online: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[42] Villavicencio, C. N., Macrohon, J. J. E., Inbaraj, X. A., Jeng, J.-H., & Hsieh, J.-G. (2021). COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. Algorithms, 14(7), 201. https://doi.org/10.3390/a14070201

[43] WHO. (2024). World Health Organization (WHO). Coronavirus 2021. Available online: https://www.who.int/health-topics/coronavirus

[44] Worldometer. (2024). Worldometer. COVID Live Update, 29 June 2021. Available online: https://www.worldometers.info/coronavirus/

[45] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., & Debray, T. P. A. (2020). Prediction Models for Diagnosis and Prognosis of COVID-19: Systematic Review and Critical Appraisal. Systematic Review and Critical Appraisal. https://doi.org/doi: 10.1136/bmj.m1328

[46] Zheng, Y., Zhu, Y., Ji, M., Wang, R., Liu, X., Zhang, M., Liu, J., Zhang, X., Qin, C. H., Fang, L., & Ma, S. (2020). A Learning-Based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics. Patterns, 1(6), 100092. https://doi.org/10.1016/j.patter.2020.100092