information

Article

Supervised Learning Models for the Preliminary Detection of COVID-19 in Patients Using Demographic and Epidemiological Parameters

Aditya Pradhan [1], Srikanth Prabhu [1,*], Krishnaraj Chadaga [1], Saptarshi Sengupta [2] and Gopal Nath [3]

[1] Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576014, India; aditya.pradhan@learner.manipal.edu (A.P.); krishnaraj.chadaga1@learner.manipal.edu (K.C.)
[2] Department of Computer Science, San Jose State University, 1 Washington Sq, San Jose, CA 95192, USA; sengupta.sap@gmail.com
[3] Department of Mathematics and Statistics, Murray State University, Murray, KY 42071, USA; gnath@murraystate.edu
* Correspondence: srikanth.prabhu@manipal.edu

Abstract: The World Health Organization labelled the new COVID-19 breakout a public health crisis of worldwide concern on 30 January 2020, and it was named the new global pandemic in March 2020. It has had catastrophic consequences on the world economy and well-being of people and has put a tremendous strain on already-scarce healthcare systems globally, particularly in underdeveloped countries. Over 11 billion vaccine doses have already been administered worldwide, and the benefits of these vaccinations will take some time to appear. Today, the only practical approach to diagnosing COVID-19 is through the RT-PCR and RAT tests, which have sometimes been known to give unreliable results. Timely diagnosis and implementation of precautionary measures will likely improve the survival outcome and decrease the fatality rates. In this study, we propose an innovative way to predict COVID-19 with the help of alternative non-clinical methods such as supervised machine learning models to identify the patients at risk based on their characteristic parameters and underlying comorbidities. Medical records of patients from Mexico admitted between 23 January 2020 and 26 March 2022, were chosen for this purpose. Among several supervised machine learning approaches tested, the XGBoost model achieved the best results with an accuracy of 92%. It is an easy, non-invasive, inexpensive, instant and accurate way of forecasting those at risk of contracting the virus. However, it is pretty early to deduce that this method can be used as an alternative in the clinical diagnosis of coronavirus cases.

Keywords: COVID-19 diagnosis; machine learning; data-driven approaches; SMOTE; SHAP; LIME; infection prediction

Citation: Pradhan, A.; Prabhu, S.; Chadaga, K.; Sengupta, S.; Nath, G. Supervised Learning Models for the Preliminary Detection of COVID-19 in Patients Using Demographic and Epidemiological Parameters. *Information* **2022**, *13*, 330. https://doi.org/10.3390/info13070330

Academic Editor: Willy Susilo

Received: 2 June 2022
Accepted: 30 June 2022
Published: 10 July 2022

1. Introduction

Coronaviruses are a family of enveloped, highly diverse, single-stranded viruses and are closely correlated to RNA viruses that infect birds and mammals [1]. They have a diameter of 60–140 nm and a genome size from 26–32 kb. When viewed under an electron microscope, they appear to look like a crown due to the glycoprotein spike-like projections on their surface, which resemble a solar corona [2]. Even though the majority of human coronaviruses (HCoV-NL63, HCoV-OC43, HCoV-229E, and HCoV-HKU1) cause minor illnesses, the epidemics of two betacoronaviruses (β-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV), in the last two decades have resulted in high mortality rates of 37% and 10%, respectively [3]. The novel coronavirus disease of 2019, also known as COVID-19, is caused by a strain of coronavirus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Its symptoms include cough, fever, fatigue, shortness of breath, body aches and loss

of taste or smell [4]. Epidemiological studies have shown that elderly individuals are more prone to severe illnesses, while children often have milder symptoms [5,6]. People with underlying severe medical conditions, such as diabetes, hypertension, cancer, lung/liver or kidney disease, have shown a bad prognosis and are at a higher risk of hospitalisation. In worst-case scenarios, the infection can be fatal [7,8]. The first case originated in Wuhan, Hubei Province, China, in December 2019 and has since spread to the entire world [9]. As of June 2022, over 532 million cases have been reported, and around 6.3 million deaths have been recorded [10].

COVID-19 is highly contagious and can transmit through direct contact (human-to-human transmission and droplet) and indirect contact (airborne contagion and contaminated objects) [11]. Its symptoms typically manifest between 1 and 14 days, while the mean incubation period is 5.2 days [12]. Countries worldwide have enforced norms such as social distancing, face masks, quarantine and vaccinations to curb the spread of this dangerous virus. Since it spreads rapidly and has no effective cure, the most efficient method of tackling its spread is early detection and isolation of patients. Currently, to diagnose COVID-19, there are two major types of tests: the first being the molecular/nucleic acid tests which include the reverse transcription-polymerase chain reaction (RT-PCR) test, digital PCR, isothermal nucleic acid amplification test and clustered regularly interspaced short palindromic repeats (CRISPR) system that detect the RNA component of the virus [13]. The RT-PCR test is considered the gold standard technique worldwide to detect COVID-19 since it delivers results more rapidly and accurately than others [14–17]. However, RT-PCR has difficulty discriminating true positives from true negatives in COVID-19 affected patients [18]. Another flaw is the false-negative rates which are highly variable. The false-negative rates are maximum during the first five days after exposure (up to 67%) and least during the eight-day after exposure (21%) [19]. Furthermore, due to the acute shortage of RT-PCR test kits in underdeveloped countries, testing and detection are delayed. The second type of test is the rapid antigen test (RAT). This test identifies antigens and small proteins on the virus's surface and gives the result within 15–30 min. Its primary disadvantages have been its low specificity (77.8%) and sensitivity (18.8%) [20]. Thus, there is an urgent requirement for a method that overcomes the pitfalls of the previous tests. One way to tackle this problem is by using artificial intelligence (AI) and machine learning (ML) to enhance clinical prediction since they recognise complex patterns in massive datasets [21]. With the advancement of machine learning, research can offer a strategic framework for developing automated, complicated and objective algorithmic tools for the analysis of multimodal and multidimensional biological/mathematical data. ML models can aid in the prediction of patients who are at a high risk of contracting COVID-19. This can prevent the spread and reduce fatalities. ML-powered prediction models combine numerous features to estimate the risk of infection and alleviate the burden on healthcare systems worldwide.

AI can be defined as a wide field of computer science concerned with developing models that can mimic human cognitive abilities. ML is a subclass of AI where the computer learns on its own by analysing historical data or experience and makes accurate predictions without being explicitly programmed. The historical data may be divided into two subsets for training and testing, among other configurations. For example, a classifier may be trained on the training dataset, where it learns about the various interesting patterns which discriminate the several existing classes. The trained model, i.e., the classifier, then predicts the classes of the testing dataset. There are four categories of machine learning. (a) Supervised machine learning methods are algorithms that learn from historical or prior datasets using labels to predict appropriate classes for unseen data (classification) or forecast future occurrences (regression) [22]. This learning style requires the presence of supervision in the form of labels in the training phase. The learning system's expected output is compared to the actual results. If discrepancies are discovered, they can be corrected by adjusting the model appropriately, usually through the employment of an optimisation algorithm that lowers the error indicative of the goodness of fit. (b) In

unsupervised learning, the input data are unclassified or unlabelled [23]. The algorithm does not specify the correct result, but it investigates the information in order that it may derive deductions from it, characterise unlabelled datasets and find meaningful patterns in it [24]. (c) Semi-supervised learning methods are those that fall between supervised and unsupervised learning models. They use both labelled and unlabelled data during the training process. This technique is used to increase the precision of learning [24]. (d) Reinforcement learning approaches use actions to engage with the learning environment to identify erroneous results [25]. The model is trained based upon the previous outcomes, and rewards and punishments exist for the predictions. Based on this principle, the model learns to maximise the rewards and minimise the penalties, thereby learning from the environment [26–28]. In addition to these approaches, it is important to highlight deep learning (DL), which is a subset of ML. The various deep learning architectures draw inspiration from and are built upon computational analogues of neurons in the human mind and aim to mimic how human beings learn. These techniques are representation-learning approaches with many layers of representation created by building simple yet non-linear components that change the representation at one level (beginning with the raw input) into a higher, increasingly abstract level [29]. Applications of DL can be found in the fields of natural language processing [30], image recognition [31], recommendation systems [32], speech recognition [33], medical diagnosis [34], etc., among others. Deep learning is extremely useful in learning complex patterns in data by means of developing tailored models that use different combinations of transformations. DL model performance scales with the amount of data, and its abstraction does not require the entire architecture to be hardcoded.

In this research, machine learning and deep learning algorithms are utilised to perform a preliminary diagnosis of COVID-19 using demographic and epidemiological parameters. These techniques can be extremely useful in geographical settings where medical resources are scarce or during pandemic peaks when demand is at its maximum, thereby putting strain on the resources. The article serves to emphasise the following contributions:

- Extensive review of background research: We perform a detailed review of recent work in the literature, which looks at various diagnostic procedures for COVID-19 using AI and ML. Emphasis is placed on articles which consider demographic and epidemiological parameters as part of their data.
- Pre-processing: The data are pre-processed to understand the most important parameters. Correlation techniques have been used to underline the most important columns in the dataset.
- Balancing: We use the Borderline-SMOTE technique to balance the data.
- Feature importance: We highlight relevant feature importance derivation techniques.
- Application of ML models: Machine learning and deep learning techniques have been used to derive insights from the data. As demonstrated below, the models tend to perform quite well for the considered data.
- Analysis of parameters: Information about the various parameters is obtained, and their effect on COVID-19 patients is studied. The results obtained are compared with state-of-the-art studies in the literature using similar data.
- Future directions: We provide an overview of some challenges faced and potential future directions to extend the work.

In this study, a labelled epidemiological dataset from various hospitals in Mexico is considered. The entire dataset in Spanish is pre-processed and balanced. Several classifiers are developed and are extensively evaluated using performance metrics such as accuracy, precision, recall, specificity and AUC. We also look at some popular techniques used in medical AI research, such as boosting and deep learning networks. The proposed models may augment efforts of detection and intervention and are ideally expected to reduce the heavy burden already faced by healthcare systems all around the world. The paper is organised as follows: Section 2 consists of similar studies that diagnose and forecast COVID-19 using machine learning. Section 3 elaborates on the dataset description, data

pre-processing, correlation analysis and some theoretical concepts related to ML. The performance metrics, model evaluation and description of results are explained in detail in Section 4. Section 5 highlights the key issues and future directions. Section 6 concludes the paper.

Motivation and Contributions

The SARS-CoV-2 virus has had devastating ramifications on human lives all across the world. Early detection of COVID-19 may increase the survivability odds of the patient, and reduce the further spread of the disease by isolating and quarantining the patients diagnosed as positive, thereby assisting in avoiding another COVID-19 wave and can mitigate the load on the healthcare professionals. Currently, there are different tests for detecting the COVID-19 strain in a patient but each of them has its respective drawbacks, such as having a high false-negative rate, delay in obtaining the results, expensive or even invasive. Our study proposes using machine learning classifiers as a technique to screen patients easily and precisely without the shortcomings faced by the current methods. There have been multiple coronavirus outbreaks in the past two decades. Our research can contribute to advancing the collective knowledge on the diagnosis of the virus and diminish the repercussions of another such outbreak in the future.

In our study, we used supervised binary classification algorithms of different categories such as the simple generalised linear logistic regression model, the lazy non-linear K-nearest neighbours' classifier, tree-based ensemble models involving bagging (random forest) and boosting (XGBoost and AdaBoost) methods and the deep learning-based artificial neural network classifier. The major objective of using a variety of classifiers was to obtain a thorough understanding of how well these algorithms comprehend the data and diagnose the patients; as each kind of classifier has its own strengths and weaknesses, this can help us to arrive at a conclusion as to which algorithm is better suited to deliver more accurate predictions to recognise if the case is positive or negative. We also analysed the results obtained to infer how each feature contributes to the outcome of the diagnosis by using SHAP and LIME techniques. Further discussion about the parameters is made from a medical perspective. Our models are supports to help researchers from both technical and medical fields.

2. Related Work

With rapid advancements made in increasing the computational power of machines and the development of new sophisticated algorithms revolutionising the big data niche, exponential progress has been seen in AI in the past two decades. In healthcare settings, accurate diagnosis and initiating treatment at the appropriate time are crucial. With broad impact encompassing the medical landscape, ML has transformed how we diagnose diseases, make predictions, analyse images, provide personalised treatment and aid patients. ML approaches have already been utilised to treat COVID-19, diabetes, pneumonia, cancer, dementia, liver failure and Parkinson's disease, amongst other ailments. They provide accurate detection and estimation results [35–40], and this has helped decrease human intervention in clinical practice.

From the start of the COVID-19 pandemic, we have seen a variety of areas where ML has been used extensively. Predicting the outbreak of COVID-19 in different countries, estimating the occurrence of the next wave and its severity, predicting mortality rates, contact tracing, detection of people not wearing facemasks or practising social distancing, developing vaccines to better understand the correlation of the underlying problems of the patient with mortality rate [41], etc., have been some of the use cases of ML. Early diagnosis of COVID-19 patients is critical to prevent the illness from progressing in an individual and from spreading to others. Research has shown that radiological imaging of the chest, such as computed tomography (CT) and X-ray, can be helpful in the early detection and treatment of COVID-19 [42]. A survey of recent literature reveals that COVID-19 mortality can be easily predicted using CT scans [43]. Narin et al. [44] were able to build a deep

convolutional neural network (CNN) model which was able to detect COVID-19 with an accuracy greater than 96% using chest X-ray scans. Ozturk et al. created a DL model named DarkCovidNet, which could detect COVID-19 accurately up to 98.08% from chest CT scan images [45]. According to these studies, these models could predict COVID-19 effectively and were as reliable as RT-PCR tests. Apart from that, they are much quicker and instantly produce results. However, these methods are invasive and need to have a radiology expert who can interpret the results, thus making the tests expensive. Furthermore, doctors do not recommend CT scans for all patients due to the radiation emitted by the machine, which can cause cancer [46]. X-rays are also prone to false-negative results [34], among other pitfalls.

Blood markers, epidemiological parameters and other demographic factors can be used for preliminary diagnosis of COVID-19. Unlike CT scans and X-rays, these facilities are available in all hospitals. The demographic parameters can be easily collected from patients. These tests can be used in parallel with RT-PCR tests. Muhammed et al. [47] used supervised ML models to predict COVID-19 using a Mexican epidemiological dataset. Eleven features were extracted for training the ML models. The dataset was obtained from the General Director of Epidemiology, who had published it on their website [48]. Five ML algorithms: decision trees, logistic regression, naïve Bayes, support vector machine and artificial neural networks (ANN) were deployed. The accuracies obtained by them were 94.99%, 94.4%, 94.36%, 92.4% and 89.2%, respectively. The article concluded that these models could be effectively deployed in hospitals. Quiroz-Juarez et al. [49] used ML to identify high risk coronavirus patients. The dataset obtained for this research was published by the Mexican Federal Government [48]. Four ML algorithms: neural networks, logistic regression, support vector machines and K-nearest neighbours (KNN) were used. The accuracies obtained were 93.5%, 92.1%, 92.5% and 89.3%, respectively. The article concluded that neural networks could easily outperform conventional machine earning algorithms. Prieto [50] used the Mexican dataset to forecast COVID-19 using ML and Bayesian approaches. Parameter estimation techniques were used in the beginning. Clinical analysis was performed later. The synthetic minority oversampling technique (SMOTE) was used in this research to balance the dataset. The author claimed that the techniques mentioned above are accurate and many false-positive and false-negative results have been eliminated. Iwendi et al. [51] used ML algorithms to diagnose COVID-19 in patients from Brazil and Mexico. Demographics, social and economic conditions, symptom reports and clinical factors were all considered. The models they developed obtained an accuracy of 93% for the Mexican dataset and 69% for the Brazilian dataset.

AI was used in early COVID-19 detection in [52]. Decision tree, Support Vector Machine and voting classifiers were used on the benchmarked dataset from Mexico. The best model obtained sensitivity, specificity and AUC of 75%, 61% and 72%, respectively. The results obtained were satisfactory according to the study. The effect of medical conditions on COVID-19 susceptibility was studied in [53]. Many COVID-19 datasets were considered for this research. The study claims that diabetes is a strong factor which links to COVID-19 mortality and that comorbidities such as hypertension and obesity are also important. Maouche et al. [54] used four ML algorithms: Multi-Layer Perceptron (MLP), decision tree, random forest and Gradient Boosting to diagnose COVID-19 using the Mexican dataset. The accuracies obtained by the models were 97.92%, 97.14%, 99.06% and 99.28%. Feature importance methods were used and the most important parameters were age, hypertension, pneumonia, diabetes and obesity.

Delgado-Gallegos et al. [55] used a decision tree model to understand the stress occupancy in healthcare professionals from Mexico. An accuracy of 94.1% was obtained by the models. Many frontline COVID-19 workers suffered from compulsive and xenophobia stress, according to the study. A random forest algorithm was used to predict the diagnosis of COVID-19 in [56]. A precision of 95% was obtained by the model. The article concluded that non-clinical diagnosis using information technology is going to play a crucial role in medical settings in the coming years. Mukherjee et al. [57] used KNN to diagnose

COVID-19 using a cloud-based Internet of Things (IoT) system. Seven COVID-19 datasets were used for this research. An ant colony optimization (ACO) algorithm was used for feature selection. Maximum accuracy of 97% was obtained by the models. The rest of the related articles are described in Table 1.

Table 1. Related works which diagnose and predict COVID-19 mortality using machine learning approaches.

Reference	Models	Accuracy	Critical Analysis/Findings
[58]	K-Means and Principal Component Analysis	-	The use of unsupervised learning in COVID-19 diagnosis. The use of principal component analysis in feature selection is also highlighted.
[59]	Naïve Bayes, Decision Tree, KNN, Support Vector Machine, Random Forest and Multi-layer perceptron	96%	The use of data mining to assist machine learning.
[60]	Logistic Regression and Support Vector Machine	72%	Accurate severity classification.
[61]	Decision Tree, Random Forest, Rotation Forest, Multi-Layer Perceptron, Naïve Bayes, KNN	87%	The use of rotation forest in diagnosing COVID-19.
[62]	Many ML models	87%	The main causes of COVID-19 deaths in Mexico were due to age, chronic diseases, bad eating habits and unnecessary contact with infected people.
[63]	Ensemble Algorithms	96%	The use of feature importance techniques such as Shapley Additive Values.
[64]	Random Forest, XGBoost, KNN and Logistic Regression	92%	The use of local interpretable model-agnostic explanations.
[65]	Ensemble Algorithms	85%	The use of SMOTETomek in data balancing.

3. Materials and Methods

3.1. Dataset Description

Mexico is one of the worst-hit countries with COVID-19, with over 5,770,000 cases and 325,000 deaths, as of 1 June 2022. It has one of the highest mortality rates of COVID-19, globally. There has also been an acute shortage of RT-PCR test kits, along with the fact that many of the tests were not conducted in an environment which could provide maximum accuracy. The dataset used in this research belongs to the official COVID-19 dataset provided by the General Directorate of Epidemiology in Mexico [66]. The data were compiled using a "sentinel model" in which 10% of patients with a viral respiratory diagnosis were tested for COVID-19 from 475 USMER (Unidades Monitoradas de Enfermedad Respiratoria Viral) hospitals to monitor viral respiratory diseases located throughout the country's health sector (ISSSTE, IMSS, SEMAR, SEDENA, etc.). The dataset used is taken from a period starting from 23 January 2020 to 26 March 2022, containing details of 15,519,390 tuples (rows) and 41 columns. Each row represents a patient record, while the columns are the various clinical, demographic and epidemiological parameters. The original dataset is collected in Spanish. It consists of the lab results for COVID-19 tests conducted in Mexico. This is an open-access dataset accessible to all users who need it to facilitate access, usage, reuse and redistribution. The details of the attributes are described in Table 2.

Table 2. List of attributes present in the Mexican COVID-19 dataset after converting to English (variable type is denoted in brackets).

Categories		Characteristics	
(1)	Medical History	1. Pneumonia (Integer) 2. Pregnancy (Integer) 3. Diabetes (Integer) 4. CPOD (Integer) 5. Asthma (Integer) 6. Autoimmune disease (Integer) 7. Hypertension (Integer)	8. Other chronic illnesses (Integer) 9. Cardiovascular disease (Integer) 10. Obesity (Integer) 11. Renal disease (Integer) 12. Tobacco (Integer) 13. Smoke (Integer)
(2)	Demographic Data	14. Address (String) 15. State (String) 16. Simple Address (String) 17. Origin (String) 18. Sector (String) 19. Gender (Integer) 20. Birth State (String)	21. Birth City (String) 22. Age (Integer) 23. Nationality (String) 24. Indigenous (Integer) 25. Migrant (Integer) 26. Original Country (String)
(3)	Medical Information	27. ID (Integer) 28. Last updated test date (Date) 29. Registration Number (String) 30. Hospital address (String) 31. Classification Final (Integer) 32. Delay (Integer) 33. Case Address (String) 34. Register Address (String)	35. Last update (Date) 36. Type of care (Integer) 37. ICU admission date (Date) 38. Date symptoms began (Date) 39. Patient Death date (Date) 40. Intubation (Integer) 41. Contact (Integer)

3.2. Data Pre-Processing

Since all the column names of the dataset had initially been in Spanish, translation to English was done before any pre-processing. Each column represented a different type of feature. The features were categorized into three major classes: (a) variables used for record-keeping, (b) variables used to store demographic details and (c) variables used to store clinical information about the patient. Elimination of all the record-keeping variables, such as the record id, date of update of the records, etc., was conducted since they were not relevant to the objectives of this research. We retained only two features among the demographic variables: sex and the patient's age. From the clinical information records of the patients, the following features were chosen: pneumonia, pregnancy, diabetes, COPD, asthma, autoimmune disease, hypertension, other chronic diseases, cardiovascular disease, obesity, renal chronic disease, tobacco and if the patient had contact with any other patient. From the 39 initial parameters, the number of features was reduced to 15. The "Classification_Final" was the target variable or result, which identified if the patient had COVID-19 or not. The values for input parameter "gender" in the original dataset was encoded to 1—female, 2—male and 99—if not specified. For males, the values were replaced from '2' to '0' to obtain a Boolean encoding. For other variables, they were encoded in a number format: 1 for "yes" and 2 for "no". All the twos were replaced with ones for better understanding. The attribute age had numerical values. The Classification_Final variable had an original encoding of the range 1–7 where '1' and '2' represented COVID-19 positive results, 3 indicated SARS-CoV-2 cases, the value '4' described an invalid case, the value '5' meant that a laboratory did not perform the testing, '6' indicated suspicious cases and '7' represented the SARS-CoV-2 negative cases. The rows which had the values '4', '5' and '6' for the Classification_Final column were dropped, and the values '1', '2' and '3' were replaced by a single value 1 to indicate that they were infected with COVID-19

and '7' which stated that COVID-19 negative cases were replaced by 0 to effectively obtain a Boolean '0/1' output from the model. After this step, the dataset had 14,307,250 rows and 16 columns. Further, exploratory data analysis (EDA) was performed on the dataset to understand the data better. Figure 1 describes the number of people who succumbed to COVID-19, the age distribution of the population which was tested and the number of males and females tested. The data types were analysed, and all the rows with missing values were deleted since the dataset was huge. The rows which had corrupt data were systematically eliminated too. The hospitals had already normalized the values of all the attributes.

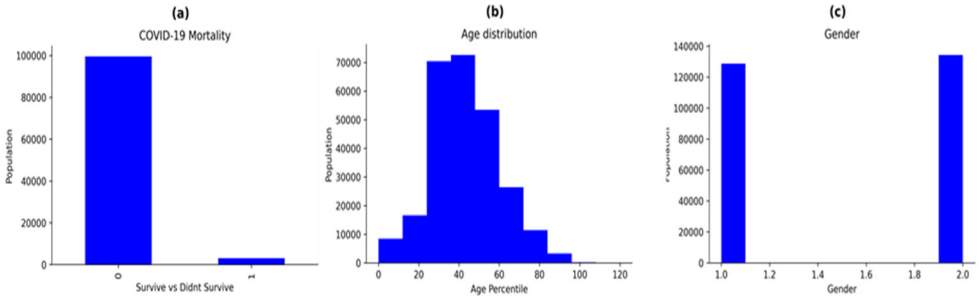

Figure 1. Statistical distribution of patients.

After initial data exploration, Pearson's correlation coefficient analysis was utilised to understand how each variable influenced the result and other variables. The Pearson correlation represented by "r" is used to understand the relationships among various parameters. If the correlation coefficient value is "1/−1" with the output, it demonstrates that there is a perfect relationship, while 0 indicates it has no effect. If the correlation coefficient value is positive, it shows that the variable affects the result positively. If it is negative, it offers an inverse impact on the output. The correlation coefficient analysis technique is based on the premise that the significance of a feature set within a dataset may be evaluated by examining the strength of the association between variables' characteristics. If the values range between 0.7 and 1.0, it is a strong correlation. If the values range between 0.3 and 0.7, it is considered a moderate correlation. Any value below 0.3 indicates a weak correlation [67]. Figure 2 shows the variables with a high and low correlation with the target variable (RT-PCR result). Some variables have a slight positive correlation relation and some variables have a slight negative correlation with the result. The "pneumonia" attribute shows the highest correlation among all variables, followed by age. This means that older adults are at an increased risk of contracting the virus. Some other interesting details from the coefficient analysis were found as well, such as men were at a higher risk of contracting the disease than women. The comorbidities also played an important role, and the features that had the most influence were hypertension, diabetes and obesity. Autoimmune diseases did not affect the result. A threshold modulus value of 0.01 was set to further eliminate the variables which had negligible influence on the output. Based on this value, the features COPD, asthma, autoimmune disease, cardiovascular disease and renal chronic disease were eliminated. This helped to narrow the dataset to the ten best features.

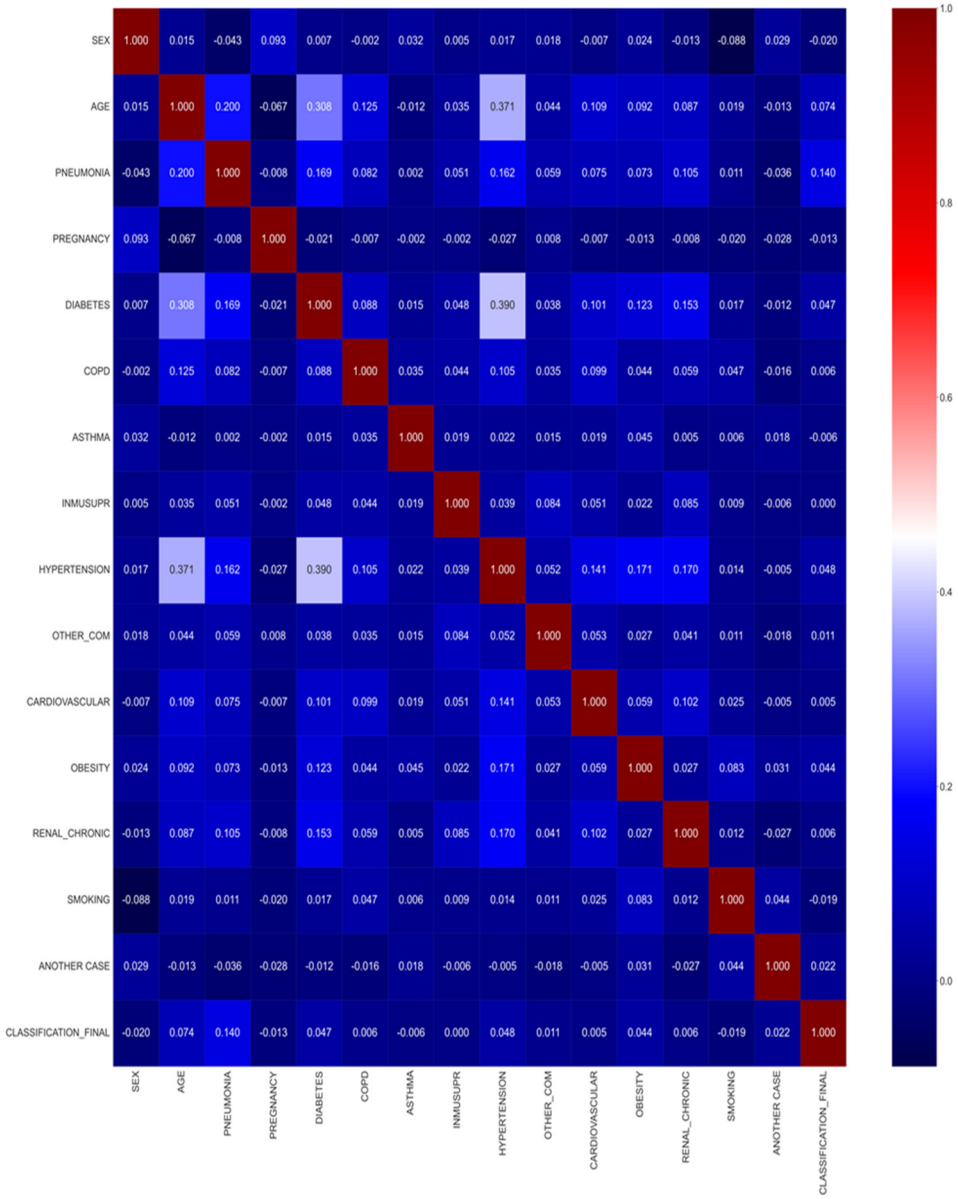

Figure 2. Pearson's correlation matrix which indicates the strength of the relationship among variables.

3.3. Some Machine Learning Algorithms and Related Terminologies

The first step in the ML process is to gather reliable data from a range of sources. This stage of data collection is critical to the modelling process. Choices such as selecting improper features or concentrating only on a subset of the data set's items might make the model less efficient. It is critical to take the required precautions while obtaining data since errors committed at this point will only exacerbate issues as advancement to the subsequent phases is made. The second step involves data preparation and processing. The primary objective of this step is to identify and mitigate any possible biases in the data sources and their characteristics. Combination of all the data and randomization of it is performed in this stage. This ensures that data are dispersed uniformly and the ordering has no effect on the learning process. Analysis of data must be done carefully to understand the data and their properties. Filtering of unnecessary features, such as names, IDs, etc., which have no significance to the model's output, were removed. Further, processing was done to find if there were any discrepancies present such as missing data, duplicate data and wrong data which can skew the results. This can be performed by visualizing the data in order to comprehend its structure and the relationships between the variables and classes.

Exploratory analysis can help us detect imbalances and relationships within the data and outliers and null values can be systematically eliminated. Further, feature scaling may be performed to have a uniform distribution of values. Data transformation by feature scaling also has other benefits such as an increased training speed, better prediction outputs and effective memory utilization. There are two major types of feature scaling: normalization and standardization. Normalization is a mapping method that creates new ranges from existing ones [68]. Of the several methods of normalization and standardization, such as (a) scaling to a range, (b) clipping, (c) log scaling and (d) Z-score, we look at min-max scaling which is a popular one where the values are converted in the range of 0 and 1 or -1 and 1.

The simple formula for min-max scalar that can scale data to a range is:

$$X' = (X - X_{min})/(X_{max} - X_{min}) \quad (1)$$

Here X, X_{max}, X_{min} and X' represent the original value, maximum value, minimum value and normalized value of the feature, respectively. Standardization, also known as variable scaling is another scaling technique [69]. It results in zero mean and unit standard deviation for each attribute in the dataset. It is also referred to as z-score normalization and can be defined as follows:

$$X' = (X - \mu)/\sigma \quad (2)$$

Here X, μ, σ and X' represent the original value, mean value, standard deviation and the standardized value for an attribute, respectively. Any of the above scaling methods can be used. To further enhance the accuracy, conversion of the string and object data type attributes to integer types is performed. There is another critical part of data processing which is segmenting the datasets into train-test splits. The bigger portion will be used to train the model, while the smaller portion will be used to evaluate it. Furthermore, the datasets should be divided in such a way that they are not leaning toward a bias. This is critical, since reusing the same datasets for training and evaluation will distort the model's efficiency. A processed input for the ML model may significantly increase its performance. It may also aid in decreasing the model's errors, resulting in increased prediction accuracy. As a result, it is essential to consider and examine the datasets to fine-tune them for better classification results. The next step is to choose a model which best aligns with the dataset. Different algorithms were created with distinct objectives in mind. It is imperative to select a model that is appropriate for the given problem from a variety of models designed for a spectrum of tasks, including voice recognition, image classification and general prediction. In this study, supervised classification algorithms were utilised to build models for predicting COVID-19 infection. Algorithms such as logistic regression, random forest, artificial neural networks (ANNs), decision trees and ensemble models such as extreme gradient boosting (XGBoost) were used for training purposes. The next step in the ML process cycle is the training stage. The pre-processed data are fed into the model which then learns the underlying patterns in it. Most of the dataset is utilized for training. This step takes a considerable amount of time as training models on large datasets with complex patterns require many iterative improvements on the part of the optimization algorithm. Once the model is trained, the final step is evaluating it to see how well it performs. It explains how well the model has been predicting by testing it on data it has not previously been exposed to, i.e., the test set. By testing it on the unseen data, we can obtain a better understanding if the model is able to adapt to new information and extrapolate to give correct outputs.

An important part of choosing the right model for the task is contingent upon successful hyperparameter tuning. Hyperparameter tuning seeks to emphasize the favourable outcomes obtained during the previous training cycles. The model is analysed and improved and this is accomplished by fine-tuning the model's parameters. The performance peaks for certain values of the parameters are retained and utilized to build the final model. The term hyperparameter tuning refers to the process of determining these values for the variables. There are several methods to determine these optimal values: one of these is to

return to the training stage and train the model using several iterations of the training data. This might result in increased accuracy since the extended length of training exposes the model to more variations of system parameters applied to the training set and increases its quality by exploring a broader region of the search space. Another approach is to refine the model's initial values. Arbitrary starting values often provide suboptimal outcomes. However, if we can improve the starting values or possibly start the model with a distribution rather than a number, the predictions may improve. There are also hyperparameters that one may tweak to observe changes in model performance. Examples of hyperparameters used in a simple model which can be altered are—learning rate, loss function applied, the training steps, etc. In this work, we use a grid search optimization technique to obtain optimized values for the parameters. Grid search is a tuning technique which performs comprehensive searching for the parameter by manually checking every value within the hyperparameter space which has been specifically defined.

Once model tuning is complete, the trained model is available for the final step in the pipeline to make predictions using the model. At this point, the model is deployed for use on unseen data. The model develops autonomy from human intervention and makes its predictions based on the test input and mapping it has learned from the training data. The machine learning algorithms used for this research are elaborated below. Figure 3 describes the process-flow of this research.

- Logistic regression: For binary and multiclass classification problems, logistic regression is an extensively used statistical classification approach. The logistic function is used to forecast the likelihood of a class label [70]. The model gives exceptional results when the labels are binary. Contrary to its name, this is a classification model, not a regression model. It is quite simple to implement and achieves excellent performance when using linearly separable classes. It uses the sigmoid function to classify the instances. The mathematical equation for logistic regression can be given as:

$$\log(P(Y)/(1-P(Y))) = \beta_0 + \beta_1 Y \tag{3}$$

  where P is the probability that Y belongs to class C and β_0 and β_1 are model parameters.
- Random forest: The random forest (RF) method is a widely used machine learning technique that interpolates the output of numerous decision trees (DT) to produce a single result [71]. It is based on the notion of ensemble learning, which is a method for integrating several weak classifiers in order to solve a complex problem. It can be used for both regression and classification problems. RF is a technique that extends the bagging approach by combining bagging with feature randomization to generate an uncorrelated forest of decision trees. It partitions the data into training and testing sets using the bootstrapping data sampling approach. The model builds trees repeatedly with each bootstrap. The final forecast is based on the average vote for each class. The larger the number of trees in the forest, the better the reliability. The chance of overfitting also decreases drastically. Further, it provides great flexibility since it can accurately perform classification and regression jobs with high accuracy. It can also be used to understand the importance of each feature. However, its main disadvantage is that these models are very complex and require much time and memory to train the models. The equations to calculate the Gini impurity and entropy are described in Equations (4) and (5). Both Gini impurity and entropy are measures of impurity of a node.

$$Gini\ Impurity = \sum_{k=1}^{c} f_k(1-f_k) \tag{4}$$

$$Entropy = \sum_{k=1}^{C} -f_i \log(f_i) \tag{5}$$

  where f is the frequency of the label and c represents the number of labels.

- XGBoost: The extreme gradient boosting (XGBoost) [72] algorithm is another prediction modelling algorithm based on ensemble learning, which can be applied to classification, regression and ranking problems. Generally, gradient boosting algorithms may suffer from overfitting as a result of data inequality [72]. However, the regularisation parameter in the XGBoost technique mitigates the danger of model overfitting. It is also an iterative tree-based ensemble classifier which seeks to improve the model's accuracy by using a boosting data resampling strategy to decrease the classification error. The algorithm is composed of a number of parameters. The ideal parameter combination improves the model's performance. It also makes use of the previous unsuccessful iteration results in the subsequent steps to achieve an optimal result. The XGBoost algorithm makes use of several CPU cores, allowing for simultaneous learning during training. The objective function of XGBoost is given by the sum of loss and regularization function as described in Equation (6).

$$obj(\theta) = \sum_{k}^{n} l(y_k - y'_{ik}) + \sum_{j=1}^{j} \omega(f_j) \tag{6}$$

  where f_j is the prediction and where j is the tree (regularisation function).
- AdaBoost: Adaptive boosting, also referred to as AdaBoost, is a machine learning approach that uses the ensemble methodology [73]. It is a meta-algorithm for statistical classification that may be used in combination with a variety of learning algorithms to enhance performance [73]. It is a widely used algorithm and it makes use of the terminology named decision stumps, which are single-level decision trees (decision trees with just one split). A key feature of AdaBoost is its adaptivity based on the results of the previous classifiers. The first step of the algorithm involves constructing a model where all data points are assigned equal weights. Points that have been misclassified are provided with larger weights. With this change, the models deployed subsequently are expected to be more reliable. The model continues to train till it reduces its loss function. However, AdaBoost's performance degrades when irrelevant features are added. It is also slow compared to XGBoost since it is not tuned for speed. The model function for AdaBoost is described in Equation (7).

$$H(x) = Sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{7}$$

The final classifier has a result $H(x)$ for x which is given by the sign of weighted summation of outcomes of T weak classifiers denoted by $h_t(x)$ and the weights assigned α_t which is calculated by using the error term of the classifier T.

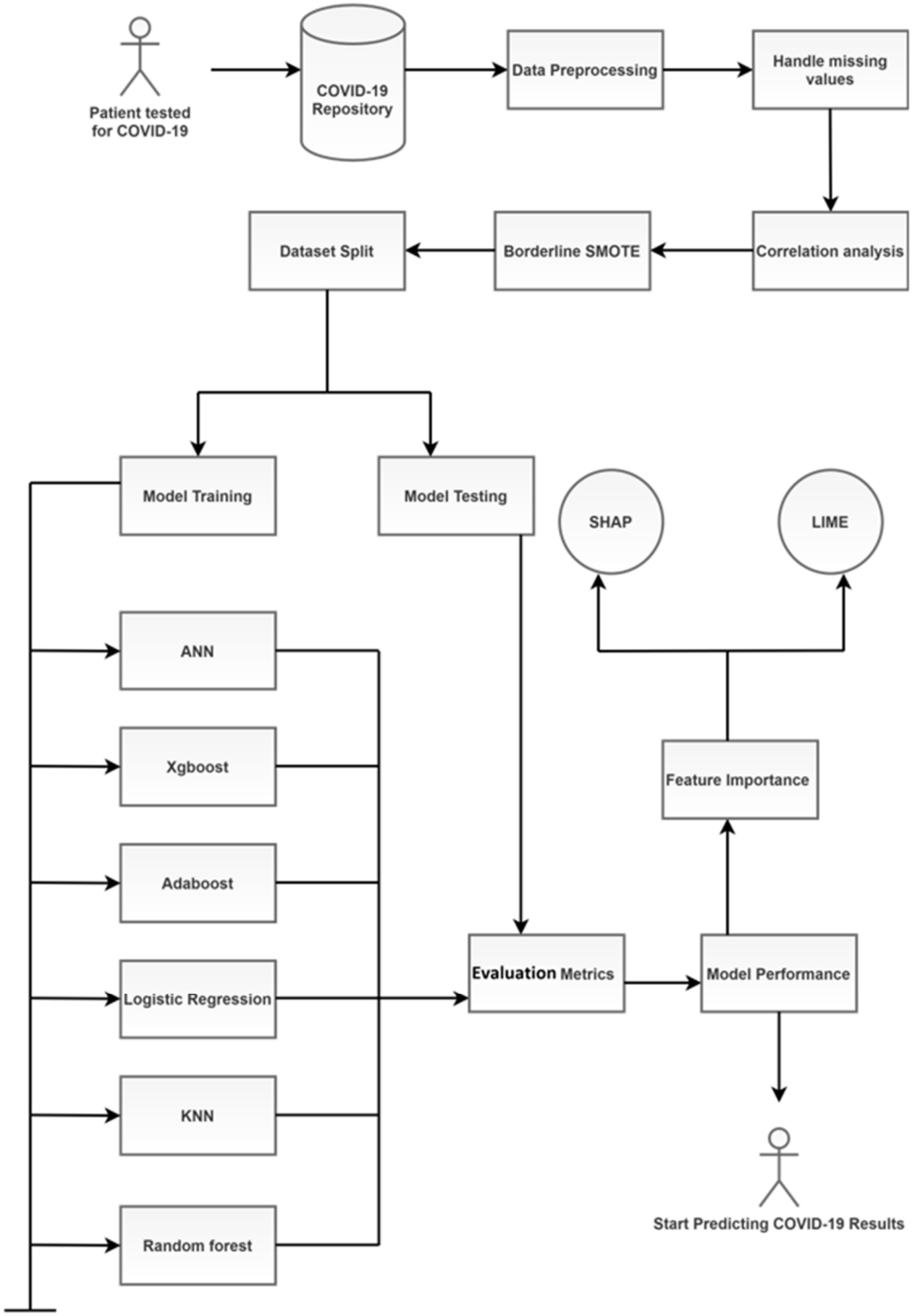

Figure 3. Process workflow of COVID-19 diagnosis using machine learning.

- KNN: The k-nearest neighbours algorithm (k-NN or KNN) is a simple non-parametric supervised ML algorithm used for both regression and classification [74]. A dataset's k closest training instances serve as the input for the model's learning process. It is also known as a "lazy learner" algorithm since it does not utilise the input during training. The KNN algorithm is based on the principle of majority voting. It gathers information from the training dataset and utilises it to make predictions about subsequent records. The first step in a KNN algorithm is to select k number of neighbours where k is an optimal constant. Calculation of the Euclidean distance (or Hamming distance for text classification) is conducted to find the nearest data points. Choosing a suitable value

of k is crucial as it affects the functioning of the algorithm. The benefits of the KNN model include its robustness, ease of implementation and its ability to pre-process large datasets. However, selecting the right k value requires expertise. Further, it also increases the computational time during testing.

- ANN: Artificial neural networks (ANNs) mirror the human brain's functioning, enabling software programs to discover patterns in large datasets [75]. They make use of nodes referred to as artificial neurons, interconnected over multiple layers of varying sizes to mimic the activities and roles of biological neural networks in the human brain. To their credit, ANNs have the ability to draw inferences about the correlations between variables which is not possible with other types of statistical models. The ANN architecture is composed of a series of node layers, they consist of a single input layer, connected to one or more hidden layers, which are then connected to an output layer. The nodes link to one another and each of them has a weight and threshold associated with it. Only when a node's output exceeds a certain threshold, is it activated and begins transferring data to the network's next layer. The node architecture for the ANN model is described in Figure 4.

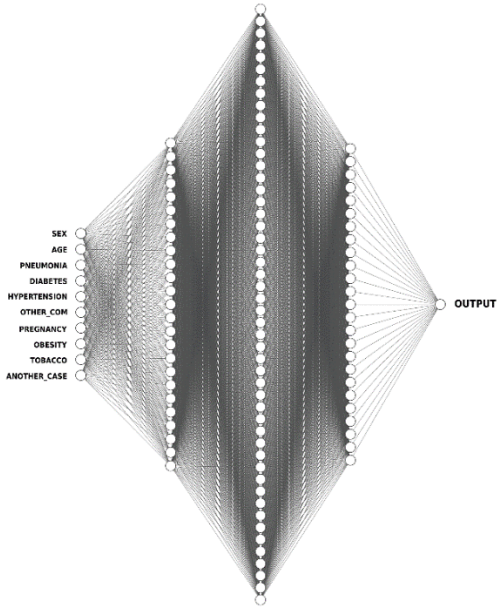

Figure 4. The architecture of ANN for this research.

- SMOTE: Data imbalance is a common problem in medical machine learning and often results in overfitting. Imbalanced class distribution has a considerable performance penalty in comparison to most traditional classifier learning techniques that assume a generally balanced class distribution and equal misclassification costs. An effective method to overcome dataset imbalance in ML is by using the synthetic minority oversampling technique (SMOTE) [76]. SMOTE employs an oversampling technique to adjust the initial training set. Rather than just replicating minority class cases, SMOTE's central concept is to offer new artificial instances which are similar to the minority class. This new dataset is constructed by interpolating between numerous occurrences of a minority class within a specific neighbourhood. In this research, a technique called the Borderline-SMOTE was used. It is based on the principle that borderline cases may provide negligible contribution to the overall success of the classification [77]. The models are more reliable when the data are balanced. Figure 5 shows the dataset before and after the use of the Borderline-SMOTE algorithm. Further, the training data were split randomly into an 80:20 ratio, with the larger proportion of the partition reserved for training the model. The smaller set was used for testing

the models' performance. It was made sure that both the subsets maintained a similar composition and lacked bias.

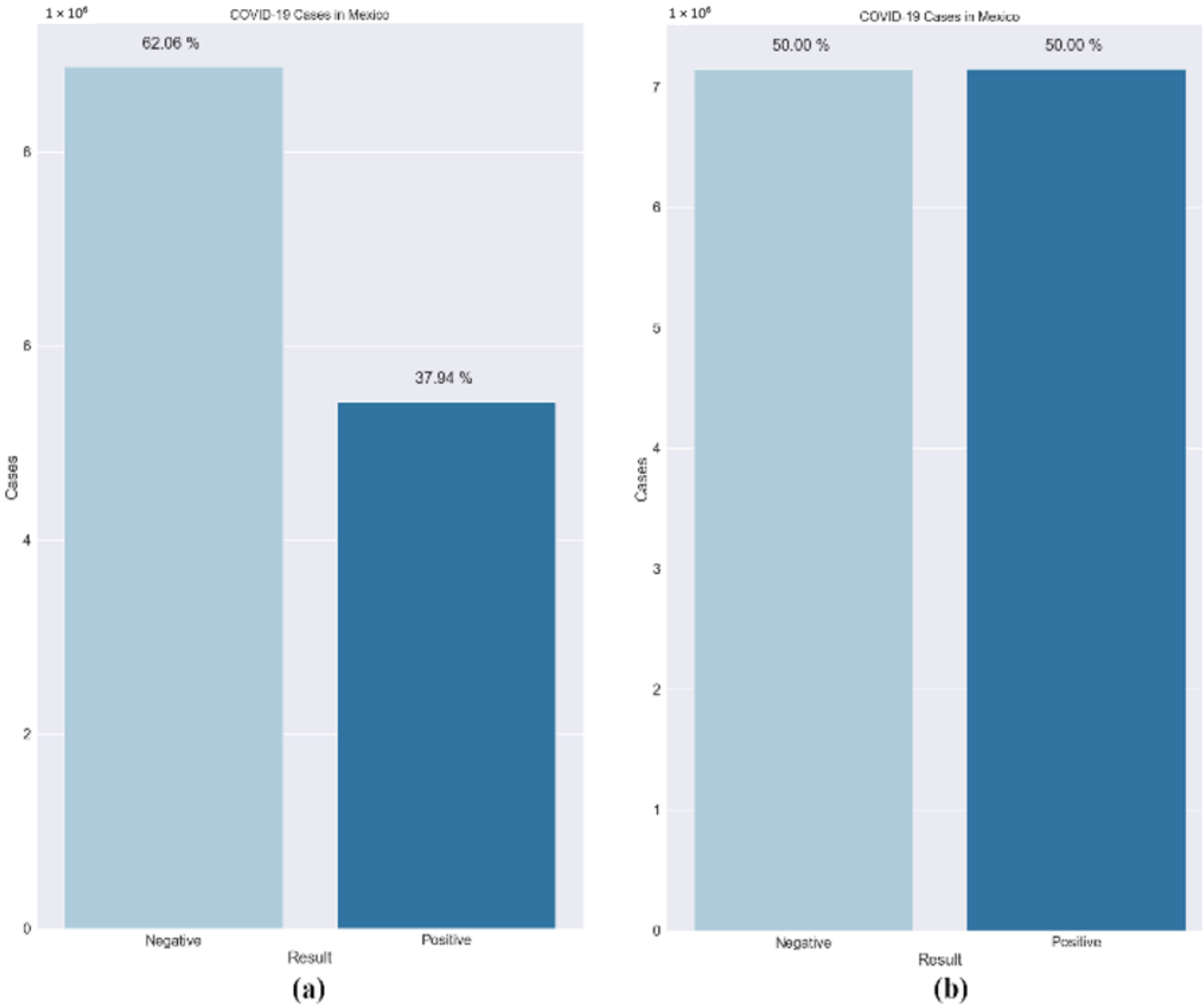

Figure 5. (**a**) Imbalanced classes, (**b**) balanced classes after using Borderline-SMOTE.

- Shapley Additive Values (SHAP): SHAP is based on the principle of game theory and it is used to increase the interpretability and transparency of the ML models [78]. Most ML and deep learning models are compatible with SHAP. The 'Tree-Explainer' procedure is mainly used in tree-based classifiers such as decision tree, random forest and other boosting algorithms. SHAP employs a variety of visual descriptions to convey the importance of attributes and how they influence the model's decision making. The baseline estimates of various parameters are compared to forecast the prediction.
- Local Interpretable Model-Agnostic Explanations (LIME): LIME is independent of any model and can be used with all the existing classifiers [79]. By adjusting the source of data points and seeing how the predictions vary, the technique seeks to understand the model's prediction. To acquire a deeper understanding of the black-box model, specific approaches look at the fundamental components and how they interact in LIME. It also modifies the attribute values in a particular order before assessing the impact on the whole outcome.

4. Results and Discussion

This study establishes a strategy for detecting COVID-19 patient outcomes by tracking patients' demographic, clinical and epidemiological characteristics. Early diagnostic forecasting of SARS-CoV-2 can help reduce the burden on the healthcare system and help save lives by predicting COVID-19 before the condition becomes extremely severe. A variety of supervised ML algorithms have been used to understand the hidden correlation between the features by utilising an epidemiological dataset of coronavirus cases in Mexico.

4.1. Performance Metrics

The model's precision, accuracy, F1-score, recall/sensitivity and AUC were all tested using the conventional assessment metrics. Additionally, a confusion matrix was also used to understand the results (true positive, true negative, false positive, false negative). The models were tested on the 20% validation data which were not used during the training phase. All classes contribute equally to the final averaged statistic in macro-average since the Borderline-SMOTE data balancing technique was used prior to training.

- Accuracy: It is a measurement which calculates the number of COVID-19 cases diagnosed accurately from the total number of cases. Correct diagnosis in this scenario is when the prediction for the case is positive, and its result is positive or when the prediction for the case is negative, and the result is also negative. It is an important metric to understand if the model is accurately diagnosing the virus. It is given by the formula:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fn + fp} \tag{8}$$

- Precision: It is another metric which calculates the ratio of patients correctly diagnosed as COVID-19 positive from the total patients predicted as COVID-19 positive by the ML models. This means that it also considers the false-positive cases, which are the patients incorrectly diagnosed with COVID-19 positive diagnosis. This metric indicates the merit of the positive cases diagnosed by the algorithm and to understand that if a patient was predicted as COVID-19 positive by the model, what would be the likelihood of them being affected by it. It is given by the formula below:

$$\text{Precision} = \frac{tp}{tp + fp} \tag{9}$$

- Recall: It is a performance metric that can be defined as the ratio of the patients correctly diagnosed as COVID-19 positive to the total patients infected by the virus. This metric emphasizes the false-negative cases. The recall is exceptionally high when the number of false-negative cases is low. It is calculated by the formula given below:

$$\text{Recall/Sensitivity} = \frac{tp}{tp + fn} \tag{10}$$

- F1-score: It is an estimate which gives equal importance to the precision and recall values obtained previously for the COVID-19 cases. It gives a better idea about the positive cases of the virus obtained. It is given by the following formula:

$$\text{F1} - \text{score} = 2 \times \frac{precision \times recall}{precision + recall} \tag{11}$$

- AUC (area under curve): The ROC (receiver operating characteristic) curve plots the true positive rate against the false-positive rate for various test instances. It indicates how well the models are differentiating the binary classes. The area under this curve is the AUC. High values for AUC indicate that the classifier is performing well.
- Confusion matrix: For binary classification, the confusion matrix is a 2×2 matrix. All the classified instances will be in the confusion matrix. The diagonal elements

indicate the correct classified instances (both true-positive and true-negative). The non-diagonal elements indicate the wrongly classified instances (both false-positive and false-negative). All the mentioned performance metrics can be easily calculated using the confusion matrix.

4.2. Model Evaluation

All the models were developed using the Python programming language in an open-source Jupyter notebook integrated development environment (IDE). All the essential libraries required for data analysis by the Python notebook, such as pandas, NumPy, Scikit, Keras, Seaborn, mathplotlib, etc., were installed and set up in the Conda virtual environment. They were used to assist in training the model and creating plots to better understand the data and results using graphical tools. They were trained on a standalone personal computer with an Intel Core i5 8th Generation processor, 16 GB RAM and 1.6 GHz processor in a Manjaro Linux operating system environment. All the models were subsequently trained with an 80:20 training–testing ratio. The confusion matrices obtained by the classifiers for the testing data are described in Figure 6. As the image depicts, the false-positive and false negative values are extremely few. This indicates that most of the COVID-19 patients' diagnoses have been predicted accurately.

XGBoost is a turbocharged decision tree-based algorithm whose strength stems from software and hardware enhancements which improve the accuracy and significantly accelerate the processes. It utilises more precise estimates to build the optimal decision tree and has been known to test the boundaries of computation by iteratively simulating every prediction depending on the error of its antecedent. During the training stage, this classifier obtained an accuracy of 94.5%. The precision, recall and F1-score values obtained were 94.7%, 93.8% and 94.2%, respectively. During the testing phase, the models obtained an accuracy, precision, recall and F1-score of 92%, 92%, 91% and 91.4%, respectively. The model was tuned to enhance its performance by modifying parameters such as the maximum depth of the tree, the learning rate and the number of trees in the ensemble model. The subsample and regularization parameters, such as alpha and lambda, were used to avoid overfitting, and for each tree, a randomised sample of columns was considered. The parameters were initially chosen by intuition and were further optimised using grid search iteratively.

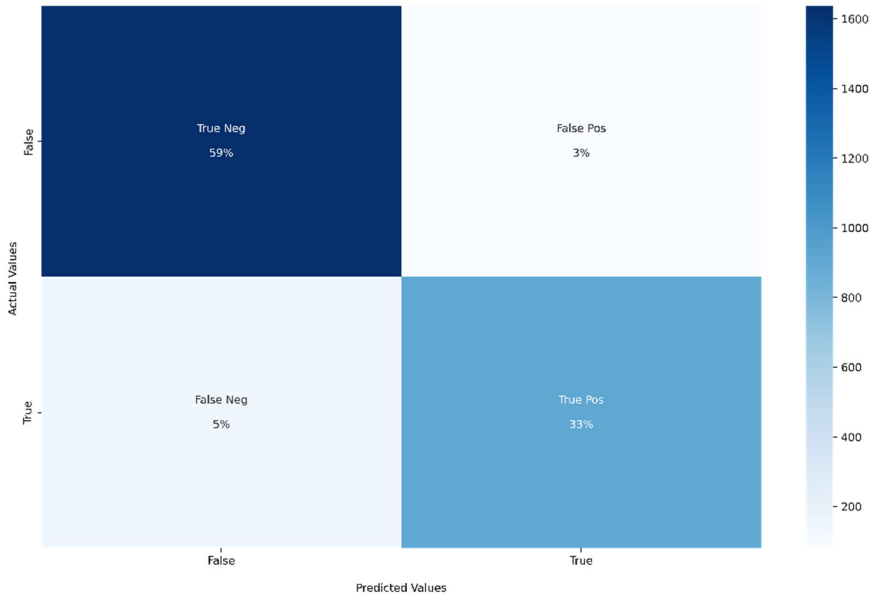

Figure 6. Confusion matrix of XGBoost algorithm.

Using AdaBoost, many flaws in the model can be improved. It gives importance to both data samples and models which makes the algorithm focus on observations which are tricky to categorize. Further, it makes use of decision stumps to sequentially train weak learners. While training, accuracy, precision, recall and F1-score values of 92.1%, 88.9%, 91.2% and 90% were achieved using the AdaBoost model. During the testing phase, the scores were 90.4%, 90.1%, 89.5% and 89.8%, respectively. The SAMME R (a new variant of the AdaBoost model) algorithm was used as it adjusts the additive model based on the probability predictions and is more accurate and quicker than the conventional classifier [80]. Apart from the above techniques, the weak learners were continuously varied using base models such as logistic regression, decision tree and random forest. Decision tree was found to be the most effective.

An artificial neural network (ANN) is made up of numerous perceptrons. Its function is to train the model by computationally mimicking, in high-level terms, the operating principles of biological neurons present in the human brain. They are constructed using several interconnected layers with weighted connections. It makes use of the concept of backpropagation to adjust weights and biases after incorporating feedback. After completion of training, it yielded accuracy, precision, recall and F1-score of 86.6%, 84.9%, 83.2% and 84.1%, respectively. For testing, it obtained accuracy, precision, recall and F1-scores of 86.2%, 88.2%, 83.1% and 85.7%, respectively. A decaying learning rate was chosen to maintain the convergence. Further, three hidden layers were used using a leaky rectified linear unit (Leaky ReLU) and sigmoid as activation functions. The adaptive moment estimation (Adam) optimizer with a batch size of 32 was utilized. ADAM is considered to be a cross between stochastic gradient descent with momentum and root mean square propagation (RMSprop) [81]. ADAM was chosen as the training cost for it was the least and it outperformed other optimisers. The number of neurons in the layers and dropouts were decided using the grid search technique.

Random forest is a collection of several decision trees. The results of the trees are combined to classify the instances based on majority voting. The first step is to create a randomised sample from the original data for each tree. For every node, a random selection of characteristics is chosen to achieve the best split possible. During training, the accuracy, precision, recall and F1-score obtained were 91%, 91.6%, 89.9% and 90.7%, respectively. During testing, the accuracy, precision, recall and F1-score obtained were 89%, 88.3%, 88.1%and 88.2%, respectively. To optimize the model's output, a variety of hyperparameter tuning methods were utilized. Tree count, node depth, the number of leaf nodes and the branch level were some of the parameters considered.

KNN assigns new data points to categories based on their similarity measure, which is often a distance measure such as Euclidean distance or Manhattan distance. It classifies new instances using a majority voting technique using the number of nearest neighbours. After training, the accuracy, precision, recall and F1-score obtained were 91.9%, 92.3%, 90.6% and 91.3%, respectively. During the testing phase, the accuracy, precision, recall and F1-score obtained were 91.6%, 91.7%, 90.5% and 91%, respectively. The most important parameter for the KNN algorithm is the value of 'K' (The number of neighbours to consider). In this research, the elbow method was used to find the optimal value of 'K' [82]. Further, the ball tree algorithm was used since the dataset was huge and had complex patterns [82]. Other parameters, such as leaf size, bias weights and metrics, were also optimized using the grid search technique.

Binary logistic regression uses the sigmoid function to classify instances. After training the model, the accuracy, precision, recall and F1-score obtained were 84.2%, 73.3%, 63.8% and 68.2%. Compared to other models, the performance of logistic regression was poor since it uses a simple approach. For testing, the model obtained accuracy, precision, recall and F1-score of 78.4%, 70%, 60.1% and 64.7%. The gradient descent algorithm was chosen with the regularization parameter 'C' whose values were tested from 0.01 to 100 for optimal hyperparameter tuning.

Table 3 summarises the results obtained by the classification algorithms. The AUCs are described in Figure 7. Experimental results demonstrated that the XGBoost model performed the best among all the classifiers. ANN, RF, AdaBoost and KNN yielded an accuracy of 86.2%, 89%, 90.4% and 91.6%, respectively. The training and testing accuracies of all the models are described in Figure 8. Further, all the metrics of all the classifiers are pictorially depicted in Figure 9.

Table 3. Summary of the results obtained by various machine learning models used in this research (in percentage).

Model	Training				Testing			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
XGBoost	94.5	94.7	93.8	94.2	92	92	91	91.4
AdaBoost	92.1	88.9	91.2	90	90.4	90.1	89.5	89.8
ANN	86.6	84.9	83.2	84.1	86.2	88.2	83.1	85.7
Random forest	91	91.6	89.9	90.7	89	88.3	88.1	88.2
KNN	91.9	92.3	90.6	91.3	91.6	91.7	90.5	91
Logistic Regression	84.2	73.3	63.8	68.2	78.4	70	60.1	64.7

XGBoost, due to enhancements in its algorithm, was able to understand the data better and give superior results. It works by enhancing the core gradient boosting machines framework using system optimisations such as pruning. The approximate greedy algorithm performs really well on the COVID-19 data because it creates trees in parallel, approximates the splits in the trees and employs its unique sparsity-aware split finding method which takes care of dense zero entities, missing values and one-hot encoded data, this is very useful for large dataset such as this one. XGBoost further takes advantages of regularization algorithms LASSO (L1) and Ridge (L2) to inflict a greater penalty on more complicated models to prevent overfitting along with its convex loss function. It also implements the quantile sketch technique to locate the ideal split locations for weighted datasets and has an inbuilt cross validation algorithm which is executed after each step. These distinctive characteristics help the XGBoost outperform the other models when they are run independently on the COVID-19 dataset.

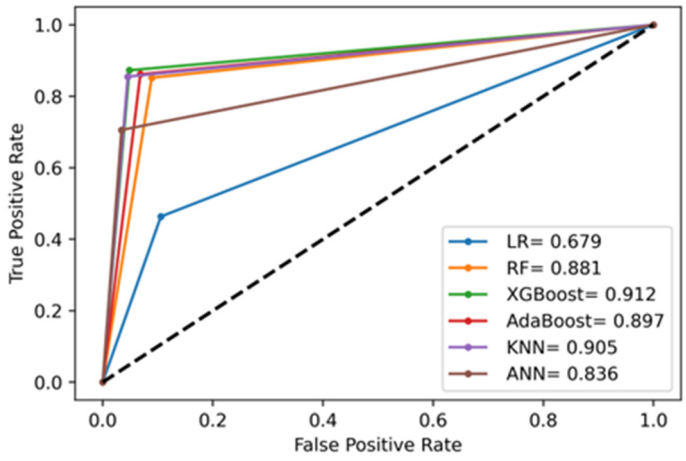

Figure 7. AUCs of various classifiers that diagnose COVID-19.

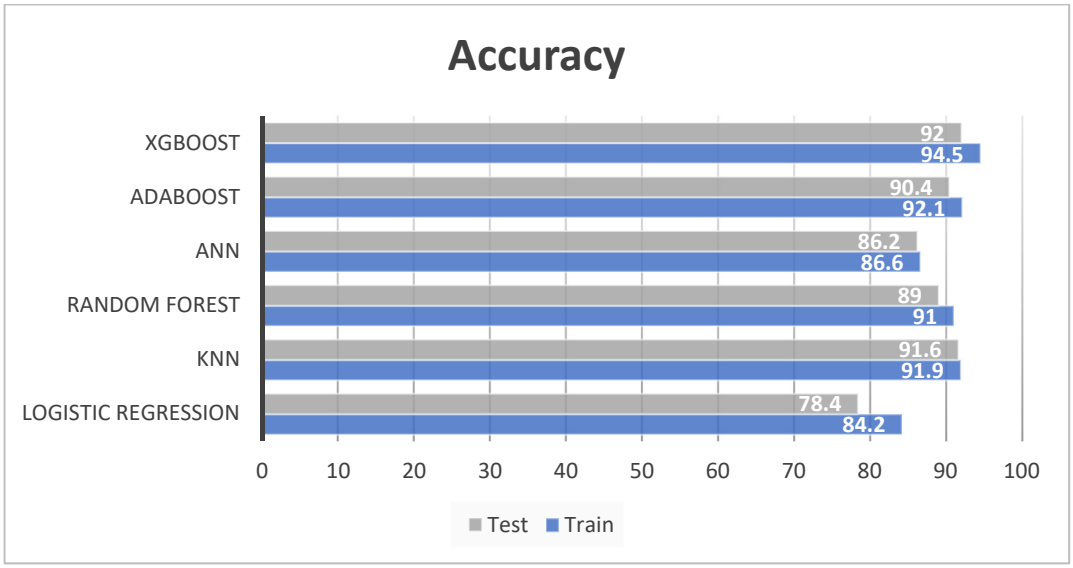

Figure 8. Training and testing accuracies of various classifiers.

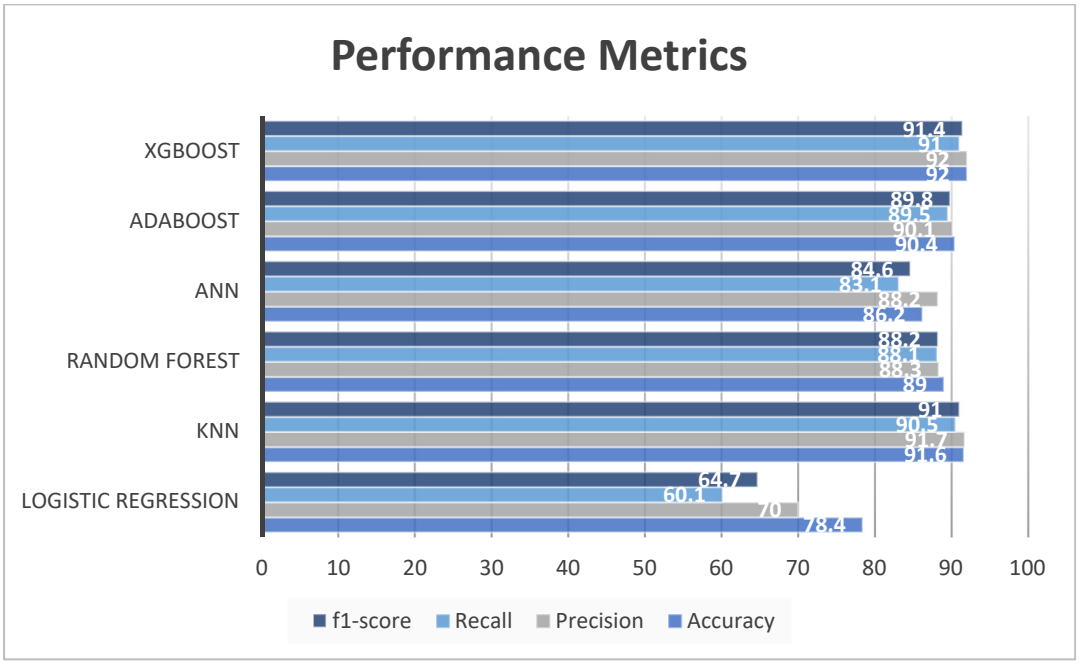

Figure 9. Performance metrics of all the classifiers during the testing phase.

RT-PCR and RAT COVID-19 testing can be supplemented using these models, which is beneficial in areas where there is an acute shortage of the above test kits. The classifiers can also be used in parallel to prevent false-negative results. It can also be highly useful during instances such as a pandemic peak. Further, these supervised ML techniques may be utilised retrospectively. This research demonstrates the potential of ML-based estimation techniques as tools augmenting interventions against the COVID-19 pandemic. With customized process pipelines in place, the described methods may also extend to enable early intervention against other diseases and new pandemics which might occur in future.

4.3. Feature Importance using SHAP and LIME

As automation becomes ever more feasible in the face of increased computational budgets, regardless of the number of ethical and legal considerations, clinical predictions derived using AI classifiers will have a tremendous impact on patient outcome going forward. Therefore, highly precise, concise and interpretable models are desirable. In the diverse medical arena, a classifier's interpretability aids the medical professional's ability to validate diagnoses made. Evaluating the algorithm's output before taking the final decision and defending treatment choices based on the classifiers are equally important. Further, feature estimates dependent on various parameters are critical for the resilience and interpretability of the models. In this research, two feature importance techniques have been utilized: (a) SHAP and (b) LIME. These two techniques help us understand the impact of various parameters in automated COVID-19 diagnosis.

SHAP understands the model using Shapley values which describe how each attribute influences the diagnosis. Figure 10a describes the bar chart obtained by SHAP. The attributes are arranged in descending order based on their importance. According to SHAP, the most important parameter is the presence of pneumonia. The other important attributes include pregnancy, sex, hypertension, age, diabetes and whether the patient was in contact with another infected patient. Further, Figure 10b explains the model precisely. The beeswarm plot also considers the value of the clinical markers. A vertical line splits the two classes. The colour "red" indicates a higher value and the colour "blue" indicates a lower value. From the figure, it can be inferred that pneumonia was mostly observed in non-COVID-19 cases (other viral infections). It can also be inferred that only a few pregnant women were susceptible to COVID-19. From a gender perspective, men are more likely to contract COVID-19. If a patient was in contact with another COVID-19 patient, there is a high percentage of chance to contract COVID-19. Older people (elderly population) are also more vulnerable in contracting this deadly virus. The presence of diabetes and obesity also plays a crucial role in diagnosing COVID-19.

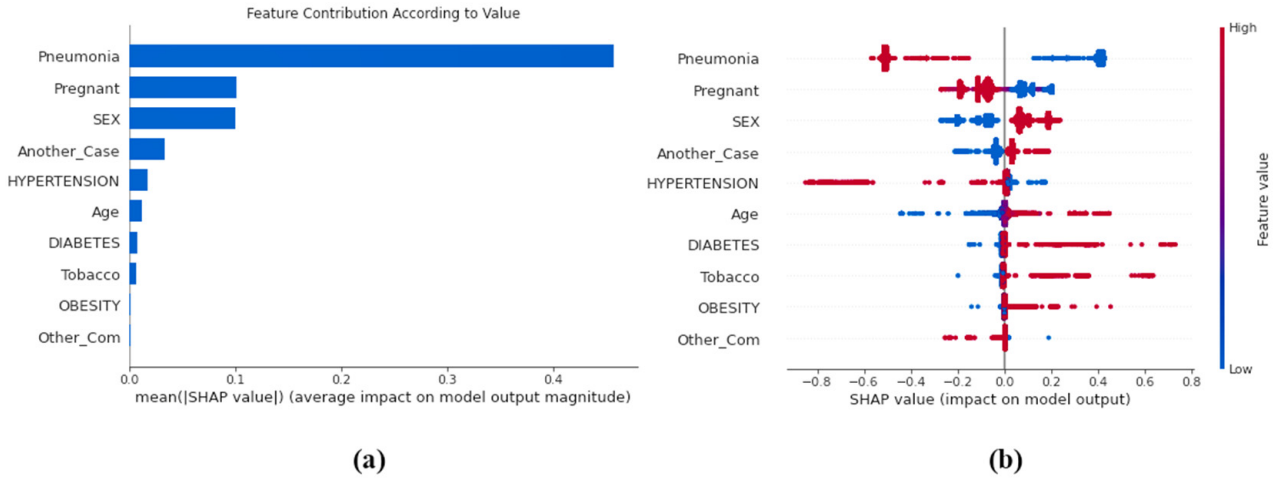

Figure 10. (**a**) Bar chart using SHAP (**b**) beeswarm plot indicating Shapley values.

LIME: The LIME feature importance models are described in Figure 11. Figure 11a describes a COVID-19 positive patient and Figure 11b describes a COVID-19 negative patient. LIME forecasts other samples by creating unique training samples near the instance to be analysed and utilizes the previous model to anticipate the cases. The instance is systematically spread based on the weights to other data points. A linear regression model is utilized based on the new samples. This approach is used to validate the learned linear model on a micro level. In Figure 11 the colour "blue" indicates COVID-19 negative diagnosis and the colour "orange" indicates a COVID-19 positive diagnosis. In Figure 11a, the prediction probability is more for the COVID-19 positive patient. The score is calculated based on various parameters such as pneumonia, age, pregnancy, diabetes and hypertension. The

weights of the parameters are also considered along with majority voting in coming to a final decision. In Figure 11b, the LIME model indicates that the patient is COVID-19 negative. All the parameters except "pregnancy" point to negative diagnosis. Using LIME, feature importance for each patient can be calculated accurately. According to explainable AI techniques, the best features obtained were pneumonia, pregnancy, sex, another_case, hypertension, age, diabetes, tobacco, obesity and other diseases.

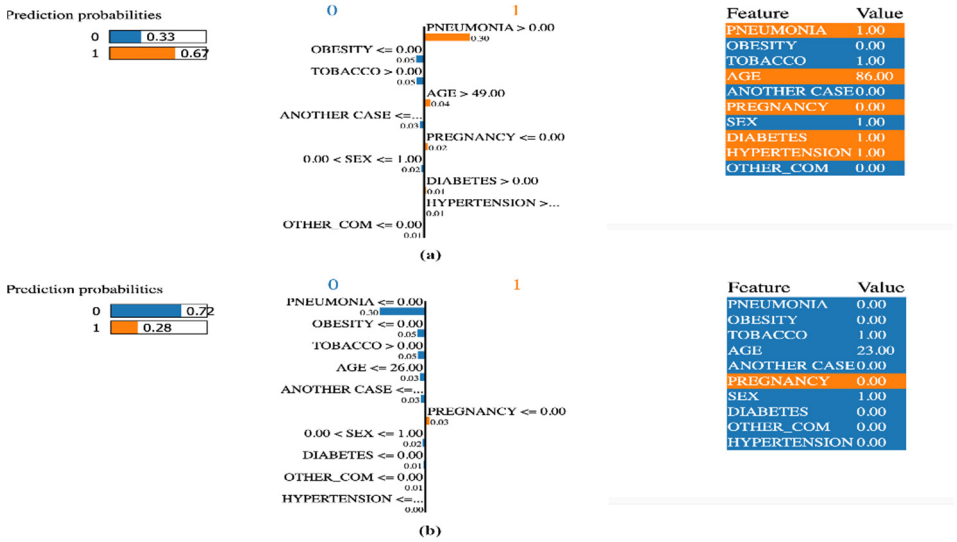

Figure 11. (**a**) COVID-19 positive diagnosis using LIME. (**b**) COVID-19 negative diagnosis using LIME.

As a retrospective evaluation technique, ML models can be deployed to predict COVID-19 diagnosis. This study describes how ML models may be built, validated and used to swiftly identify patients. The study also highlights the use of feature importance methods in identifying the most important markers. This aids in reducing the substantial workload placed on front-line health professionals. This also helps underdeveloped countries which lack technical and clinical resources under the burden of case volume during an infection peak.

4.4. Further Discussion

In this research, a set of epidemiological and demographic parameters strongly associated with COVID-19 were identified. The data also contained details of patients who had similar symptoms but were diagnosed as COVID-19 negative. Before the actual test results are obtained, these traits may help the doctors in identifying potential patients.

Many viral diseases cause pneumonia. This condition is extremely dangerous and can lead to fatality. In severe cases, COVID-19 is known to induce pneumonia along with conditions such as acute respiratory distress syndrome (ARDS) and multi-organ failure. However, in this dataset, most of the COVID-19 patients did not suffer from pneumonia. COVID-19 is known to spread among all humans including pregnant women. However, most of the pregnant women in this dataset were diagnosed as COVID-19 negative. This dangerous disease in known to spread rapidly. Nationwide lockdowns were imposed to prevent the spread of this disease. It was likely that a patient could contract COVID-19 when he was in contact with another infected patient. Patients with comorbidities, such as hypertension and diabetes, are more vulnerable to succumb to COVID-19. This research reinforces that diabetes, tobacco use and obesity increases the chance of infection. According to the study, the presence of other diseases apart from the ones mentioned above, are not extremely dangerous from an infection standpoint. Furthermore, most patients suffering from hypertension were COVID-19 negative. These are some of the main inferences made from the study.

The pandemic's heavy toll on human health and well-being has spurred various research labs to develop intelligent systems with the purpose of automating COVID-19 detection and severity. However, only a few ML models based on demographic and epidemiological models have been deployed. Muhammed et al. [48] used ML models to diagnose COVID-19 for the Mexican dataset. Five ML models were utilized and a maximum accuracy of 95% was obtained by the decision tree model. However, no feature importance techniques were utilized to understand the model's predictions. Juárez et al. [49] used the Mexican dataset for COVID-19 diagnosis. Among the four ML modes, neural network obtained the maximum accuracy of 93.5%. Iwendi et al. [51] used AI to diagnose COVID-19 for the Brazilian and Mexican patients. However, the accuracy obtained for the Mexican dataset was only 69%. Martinez-Velaquez et al. [52] used ML for early detection of COVID-19 where 22 features were considered and a maximum sensitivity of 75% was obtained. Rezapour and Colin [53] used ML to understand the relationship between COVID-19 susceptibility and comorbidities. The abovementioned works are summarized and compared in Table 4.

Table 4. Comparison of various researches in diagnosing COVID-19.

Reference	Dataset Origin	ML Models Used	No of Parameters Considered	Accuracy	Feature Importance
[48]	Mexico	Five	10	94.99%	No
[49]	Mexico	Various ML models	21	93.50%	No
[51]	Mexico	Various ML models	-	69%	No
[52]	Mexico	Various ML models	22	Sensitivity-75%	Gini Index
[53]	Mexico	Various ML models	14	Qualitative	No
Proposed	Mexico	Six	10	94.50%	SHAP and LIME

In this research, ML was used to analyse the epidemiological and demographic parameters in predicting the occurrence of infection with coronavirus causing COVID-19. These results are often easily available in shorter time intervals and at a lower price than radiographic and molecular tests. A dataset from Mexico was utilized and six machine learning models commonly used in medical AI were deployed. Information about patients, data security, integration of data and automation are the advantages of using EMR (electronic health records). We emphasize the use of data-driven models which aim to help clinicians make better decisions by providing them with valuable information generated by the trained models. Further, feature importance techniques, such as SHAP and LIME, have been effectively utilized which make the model more precise, interpretable and accurate. This helps medical professionals during the final diagnosis of the patient.

5. Challenges and Future Directions

Various challenges and clear directions for upcoming ML enthusiasts and medical professionals are provided in this section.

5.1. Challenges

With AI making progress in leaps and bounds in the development of new algorithms, it has increased the scope to where it can be applied. ML has many potential applications across different medical problems. However, there is a clear dearth of such procedures being effectively used in clinical practice. The following are some challenges that should be addressed before widespread adoption is likely.

- Data from a single country: For this research, data were collected from Mexico. However, data from all geographic areas must be considered for better validation. This is not a trivial task as there are clear differences in reporting standards and authenticity across different countries.

- Imbalance in data: In much of medical AI research, data imbalance is a persistent issue. The number of healthy patients is always more than the number of infected people. However, the models perform well when there are an equal number of classes. In this research, the Borderline-SMOTE technique was used to balance the data. Appropriate pre-processing should precede model training when working with such data.
- Original values: The data obtained for this research was already normalized. However, original data are required to form accurate medical intuitions.
- Missing blood and clinical markers: Clinical markers, such as CRP (C reactive protein), D-dimer, ferritin and lactate dehydrogenase (LDH) are known to be extremely useful in diagnosing COVID-19. However, these markers were not available in the dataset.
- Variance in computer equipment: There is no one single uniform standard architecture followed by machines universally. The data are quite sensitive to software and hardware changes of the setup.
- Distributional shift in test data: An ML model will struggle to perform well if it is unable to adapt to novel scenarios. Trained models in supervised learning are notoriously bad at detecting meaningful changes in context or data, which leads to inaccurate predictions based on out-of-scope data. When the ML method is incorrectly applied to an unexpected patient situation, it might cause a disparity between the learning and operational data.
- Difficulties in deploying AI systems on a logistical level: Numerous existing difficulties in converting AI applications to clinical practice are due to the fact that the majority of healthcare data are not easily accessible for machine learning. Data are often compartmentalised in a plethora of medical imaging archiving systems, electronic health records (EHR), pathology systems, electronic prescription tools and insurance databases, making integration very challenging.
- Interpreting the result: The model may be able to derive complex and hidden patterns. However, sometimes these patterns might have no meaning. This might be problematic in medical applications, where there is a high need for techniques that are not just effective, but also clear, interpretable and explainable.
- Quality of data: It is essential to obtain reliable input from authentic sources. It is also necessary to filter out the noise which may have crept in while feeding the data.
- Data privacy: Most of the medical data obtained from the patients are highly confidential. A leak, attack or misuse of it can be catastrophic.

5.2. Future Directions

- Improving the dataset: For further research, a more balanced dataset can be collected. Important clinical markers mentioned in the previous section can also be considered. COVID-19 severity can also be predicted.
- Using different algorithms: This research can be expanded by experimenting with different ML algorithms and combining them, as each model has its own pros and cons, there could be a model which is tailor-made for this dataset
- Medical validation: Medical validation can be performed by doctors to comment on the authenticity of the models. Further, the models can be deployed in medical facilities and feedback on accuracy can be incorporated.
- Combining other AI methodologies: CT-scans, X-rays, MRIs, ultrasound and cough sound analysis also use AI to diagnose COVID-19. The integration of these models is expected to produce compelling results.

6. Conclusions

COVID-19 must be diagnosed as early as possible for the patients to obtain appropriate treatment and prevent it from spreading to others. In recent studies, it has been proved that laboratory markers are an excellent diagnosis method since they are relatively cheap and easily available in most hospitals for implementation schemes using data-driven techniques. In this work, an extensive review of related literature was conducted in the beginning. The

dataset used in this research contained epidemiological and demographic characteristics of patients from Mexico who were tested for COVID-19. Data pre-processing was performed subsequently, followed by correlation analysis. The 10 best features were chosen for training the ML models. Six popular ML classifiers commonly used in medical AI were trained and tested. Among all the models, XGBoost achieved the highest accuracy with 94.5% during training and 92% while testing. To understand the importance of each attribute, feature importance methods, such as SHAP and LIME, were utilized. Furthermore, the proposed models were compared with the other state-of-the-art models and the reliability and effectiveness of the tested models were determined.

There is much scope for improvement in automated COVID-19 diagnosis. For accurate and precise predictions, various factors have to be addressed, particularly in modern clinical settings. Good quality data, rigorous testing and external validation must be conducted by ML and medical researchers in the near future. The trained models aim at realizing a relatively easy and inexpensive mode of quick detection of cases that may lessen the burden on healthcare workers by augmenting their efforts, especially during periods of increased caseload.

Author Contributions: Conceptualization, K.C.; methodology, K.C., S.P. and A.P.; validation, S.S.; formal analysis, S.P.; writing—original draft preparation, A.P.; writing—review and editing, S.S., K.C. and S.P.; visualization, S.S.; supervision, G.N. and K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Woo, P.C.; Huang, Y.; Lau, S.K.; Yuen, K.Y. Coronavirus genomics and bioinformatics analysis. *Viruses* **2010**, *2*, 1804–1820. [CrossRef] [PubMed]
2. Hayden, F.; Richman, D.; Whitley, R. *Clinical Virology*, 4th ed.; ASM Press: Washington, DC, USA, 2017.
3. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef]
4. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **2020**, *5*, 536–544. [CrossRef] [PubMed]
5. Yuki, K.; Fujiogi, M.; Koutsogiannaki, S. COVID-19 pathophysiology: A review. *Clin. Immunol.* **2020**, *215*, 108427. [CrossRef]
6. Liu, K.; Chen, Y.; Lin, R.; Han, K. Review-Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *J. Infect.* **2020**, *80*, e14–e18. [CrossRef]
7. Singh, A.K.; Gupta, R.; Ghosh, A.; Misra, A. Diabetes in COVID-19: Prevalence, pathophysiology, prognosis and practical considerations. *Diabetes Metab. Syndr.* **2020**, *14*, 303–310. [CrossRef]
8. Zhang, J.; Wang, X.; Jia, X.; Li, J.; Hu, K.; Chen, G.; Wei, J.; Gong, Z.; Zhou, C.; Yu, H.; et al. Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China. *Clin. Microbiol. Infect.* **2020**, *26*, 767–772. [CrossRef]
9. Lu, H.; Stratton, C.W.; Tang, Y.W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *J. Med. Virol.* **2020**, *92*, 401–402. [CrossRef]
10. Johns Hopkins Coronavirus Resource Center. Available online: https://coronavirus.jhu.edu/ (accessed on 1 June 2022).
11. Lei, S.; Jiang, F.; Su, W.; Chen, C.; Chen, J.; Mei, W.; Zhan, L.; Jia, Y.; Zhang, L.; Liu, D.; et al. Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection. *EClinicalMedicine* **2020**, *21*, 100331. [CrossRef]
12. Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K.; Lau, E.; Wong, J.; et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.* **2020**, *382*, 1199–1207. [CrossRef]
13. Habibzadeh, P.; Mofatteh, M.; Silawi, M.; Ghavami, S.; Faghihi, M. Molecular diagnostic assays for COVID-19: An overview. *Crit. Rev. Clin. Lab. Sci.* **2021**, *58*, 385–398. [CrossRef] [PubMed]

14. Mahendiratta, S.; Batra, G.; Sarma, P.; Kumar, H.; Bansal, S.; Kumar, S.; Prakash, A.; Sehgal, R.; Medhi, B. Molecular diagnosis of COVID-19 in different biologic matrix, their diagnostic validity and clinical relevance: A systematic review. *Life Sci.* **2020**, *258*, 118207. [CrossRef] [PubMed]
15. Goudouris, E.S. Laboratory diagnosis of COVID-19. *J. Pediatr.* **2021**, *97*, 7–12. [CrossRef] [PubMed]
16. Zhu, H.; Zhang, H.; Xu, Y.; Laššáková, S.; Korabečná, M.; Neužil, P. PCR past, present and future. *BioTechniques* **2020**, *69*, 317–325. [CrossRef]
17. Falzone, L.; Gattuso, G.; Tsatsakis, A.; Spandidos, D.A.; Libra, M. Current and innovative methods for the diagnosis of COVID-19 infection (Review). *Int. J. Mol. Med.* **2021**, *47*, 100. [CrossRef]
18. Yang, Y.; Yang, M.; Yuan, J.; Wang, F.; Wang, Z.; Li, J.; Zhang, M.; Xing, L.; Wei, J.; Peng, L.; et al. Laboratory Diagnosis and Monitoring the Viral Shedding of SARS-CoV-2 Infection. *Innovation* **2020**, *1*, 100061. [CrossRef]
19. Kucirka, L.M.; Lauer, S.A.; Laeyendecker, O.; Boon, D.; Lessler, J. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. *Ann. Intern. Med.* **2020**, *173*, 262–267. [CrossRef]
20. Burog, A.; Yacapin, C.; Maglente, R.; Macalalad-Josue, A.; Uy, E.; Dans, A.; Dans, L. Should IgM/IgG rapid test kit be used in the diagnosis of COVID-19? *Acta Med. Philipp.* **2020**, *54*, 1–12. [CrossRef]
21. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef]
22. Rustam, F.; Reshi, A.A.; Mehmood, A.; Ullah, S.; On, B.; Aslam, W.; Choi, G.S. COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access* **2020**, *8*, 101489–101499. [CrossRef]
23. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **2007**, *160*, 3–24.
24. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
25. Liu, D.; Clemente, L.; Poirier, C.; Ding, X.; Chinazzi, M.; Davis, J.T.; Vespignani, A.; Santillana, M. A machine learning methodology for real-time forecasting of the 2019–2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv* **2020**, arXiv:2004.04019.
26. Saravanan, R.; Sujatha, P. A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In Proceedings of the IEEE 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 14–15 June 2018; pp. 945–949.
27. Kaelbling, L.; Littman, M.; Moore, A. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
28. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
29. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
30. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
31. Pak, M.S.; Kim, S.H. A review of deep learning in image recognition. In Proceedings of the International Conference on Computer Applications and Information Processing Technology, Kuta Bali, Indonesia, 8–10 August 2017; pp. 1–3.
32. Shokeen, J.; Rana, C. An Application-oriented Review of Deep Learning in Recommender Systems. *Int. J. Intell. Syst. Appl.* **2019**, *11*, 46–54. [CrossRef]
33. Lee, W.; Seong, J.J.; Ozlu, B.; Shim, B.S.; Marakhimov, A.; Lee, S. Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review. *Sensors* **2021**, *21*, 1399. [CrossRef]
34. Chadaga, K.; Prabhu, S.; Vivekananda, B.K.; Niranjana, S.; Umakanth, S. Battling COVID-19 using machine learning: A review. *Cogent Eng.* **2021**, *8*, 1958666. [CrossRef]
35. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* **2018**, *9*, 515. [CrossRef]
36. Toğaçar, M.; Ergen, B.; Cömert, Z.; Özyurt, F. A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models. *IRBM* **2020**, *41*, 212–222. [CrossRef]
37. Kourou, K.; Exarchos, T.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef] [PubMed]
38. Pellegrini, E.; Ballerini, L.; Hernandez, M.D.C.V.; Chappell, F.M.; González-Castro, V.; Anblagan, D.; Danso, S.; Muñoz-Maniega, S.; Job, D.; Pernet, C.; et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer Dement. Diagn. Assess. Dis. Monit.* **2018**, *10*, 519–535. [CrossRef]
39. Bind, S.; Tiwari, A.K.; Sahani, A.K. A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction. *Int. J. Comput. Sci. Inf. Technol.* **2015**, *6*, 1648–1655.
40. Musunuri, B.; Shetty, S.; Shetty, D.K.; Vanahalli, M.K.; Pradhan, A.; Naik, N.; Paul, R. Acute-on-Chronic Liver Failure Mortality Prediction using an Artificial Neural Network. *Eng. Sci.* **2021**, *15*, 187–196. [CrossRef]
41. Lalmuanawma, S.; Hussain, J.; Chhakchhuak, L. Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review. *Chaossolitons Fractals* **2020**, *139*, 110059. [CrossRef]
42. Zu, Z.Y.; Jiang, M.D.; Xu, P.P.; Chen, W.; Ni, Q.Q.; Lu, G.M.; Zhang, L.J. Coronavirus Disease 2019 (COVID-19): A Perspective from China. *Radiology* **2020**, *296*, E15–E25. [CrossRef]
43. Lee, E.Y.P.; Ng, M.-Y.; Khong, P.-L. COVID-19 pneumonia: What has CT taught us? *Lancet Infect. Dis.* **2020**, *20*, 384–385. [CrossRef]
44. Narin, A.; Kaya, C.; Pamuk, Z. Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. *Pattern Anal. Appl.* **2021**, *24*, 1207–1220. [CrossRef]

45. Ozturk, T.; Talo, M.; Yildirim, E.A.; Baloglu, U.B.; Yildirim, O.; Acharya, U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **2020**, *121*, 103792. [CrossRef]
46. Smith-Bindman, R.; Yu, S.; Wang, Y.; Kohli, M.D.; Chu, P.; Chung, R.; Luong, J.; Bos, D.; Stewart, C.; Bista, B.; et al. An Image Quality–informed Framework for CT Characterization. *Radiology* **2022**, *302*, 380–389. [CrossRef]
47. Muhammad, L.J.; Algehyne, E.A.; Usman, S.S.; Ahmad, A.; Chakraborty, C.; Mohammed, I.A. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Comput. Sci.* **2020**, *2*, 11. [CrossRef] [PubMed]
48. Franklin, M.R. Mexico COVID-19 Clinical Data. Available online: https://www.kaggle.com/marianarfranklin/mexico-covid19-clinical-data/metadata (accessed on 26 June 2020).
49. Quiroz-Juárez, M.A.; Torres-Gómez, A.; Hoyo-Ulloa, I.; León-Montiel, R.D.J.; U'Ren, A.B. Identification of high-risk COVID-19 patients using machine learning. *PLoS ONE* **2021**, *16*, e0257234. [CrossRef] [PubMed]
50. Prieto, K. Current forecast of COVID-19 in Mexico: A Bayesian and machine learning approaches. *PLoS ONE* **2022**, *17*, e0259958. [CrossRef] [PubMed]
51. Iwendi, C.; Huescas, C.; Chakraborty, C.G.Y.; Mohan, S. COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients. *J. Exp. Theor. Artif. Intell.* **2022**, *1*, 1–21. [CrossRef]
52. Martinez-Velazquez, R.; Tobon, V.D.P.; Sanchez, A.; El Saddik, A.; Petriu, E. A Machine Learning Approach as an Aid for Early COVID-19 Detection. *Sensors* **2021**, *21*, 4202. [CrossRef]
53. Rezapour, M.; Varady, C.A. A machine learning analysis of the relationship between some underlying medical conditions and COVID-19 susceptibility. *arXiv* **2021**, arXiv:2112.12901.
54. Maouche, I.; Terrissa, S.L.; Benmohammed, K.; Zerhouni, N.; Boudaira, S. Early Prediction of ICU Admission Within COVID-19 Patients Using Machine Learning Techniques. In *Innovations in Smart Cities Applications*; Springer: Cham, Switzerland, 2021; Volume 5, pp. 507–517.
55. Delgado-Gallegos, J.L.; Avilés-Rodriguez, G.; Padilla-Rivas, G.R.; Cosio-León, M.D.l.Á.; Franco-Villareal, H.; Zuñiga-Violante, E.; Romo-Cardenas, G.S.; Islas, J.F. Clinical applications of machine learning on COVID-19: The use of a decision tree algorithm for the assessement of perceived stress in mexican healthcare professionals. *medRxiv* **2020**. [CrossRef]
56. Yadav, A. Predicting Covid-19 using Random Forest Machine Learning Algorithm. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Khargpur, India, 6 July 2021; pp. 1–6.
57. Mukherjee, R.; Kundu, A.; Mukherjee, I.; Gupta, D.; Tiwari, P.; Khanna, A.; Shorfuzzaman, M. IoT-cloud based healthcare model for COVID-19 detection: An enhanced k-Nearest Neighbour classifier based approach. *Computing* **2021**, 1–21. [CrossRef]
58. Chaudhary, L.; Singh, B. Community detection using unsupervised machine learning techniques on COVID-19 dataset. *Soc. Netw. Anal. Min.* **2021**, *11*, 28. [CrossRef]
59. Cornelius, E.; Akman, O.; Hrozencik, D. COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty. *Mathematics* **2021**, *9*, 2043. [CrossRef]
60. Wollenstein-Betech, S.; Cassandras, C.G.; Paschalidis, I.C. Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for and ICU or ventilator. *Int. J. Med. Inform.* **2020**, *123*, 11–22. [CrossRef] [PubMed]
61. Durden, B.; Shulman, M.; Reynolds, A.; Phillips, T.; Moore, D.; Andrews, I.; Pouriyeh, S. Using Machine Learning Techniques to Predict RT-PCR Results for COVID-19 Patients. In Proceedings of the 2021 IEEE Symposium on Computers and Communications (ISCC), Athens, Greece, 5–8 September 2021; pp. 1–4.
62. Guzmán-Torres, J.A.; Alonso-Guzmán, E.M.; Domínguez-Mota, F.J.; Tinoco-Guerrero, G. *Estimation of the Main Conditions in (SARS-CoV-2) COVID-19 Patients That Increase the Risk of Death Using Machine Learning, the Case of Mexico*; Elsevier: Amsterdam, The Netherlands, 2021; Volume 27.
63. Chadaga, K.; Prabhu, S.; Umakanth, S.; Bhat, V.K.; Sampathila, N.; Chadaga, R.P.; Prakasha, K.K. COVID-19 Mortality Prediction among Patients Using Epidemiological Parameters: An Ensemble Machine Learning Approach. *Eng. Sci.* **2021**, *16*, 221–233. [CrossRef]
64. Chadaga, K.; Chakraborty, C.; Prabhu, S.; Umakanth, S.; Bhat, V.; Sampathila, N. Clinical and laboratory approach to diagnose COVID-19 using machine learning. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14*, 452–470. [CrossRef] [PubMed]
65. Almansoor, M.; Hewahi, N.M. Exploring the Relation between Blood Tests and COVID-19 Using Machine Learning. In Proceedings of the 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 26–27 October 2020; pp. 1–6.
66. Open Data General Directorate of Epidemiology. Available online: https://www.gob.mx/salud/documentos/datos-abiertos-152127 (accessed on 26 March 2022).
67. Ahlgren, P.; Jarneving, B.; Rousseau, R. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 550–560. [CrossRef]
68. Devillanova, G.; Solimini, S. Min-max solutions to some scalar field equations. *Adv. Nonlinear Stud.* **2012**, *12*, 173–186. [CrossRef]
69. Thara, T.D.K.; Prema, P.S.; Xiong, F. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognit. Lett.* **2019**, *128*, 544–550.
70. Nick, T.G.; Campbell, K.M. Logistic regression. *Methods Mol. Biol.* **2007**, *404*, 273–301.
71. Belgiu, M.; Drăguţ, L. Random Forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]

72. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
73. Schapire, R.E. Explaining adaboost. In *Empirical Inference*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 37–52.
74. Zhang, M.; Zhou, Z. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
75. Krogh, A. What are Artificial Neural Networks? *Nat. Biotechnol.* **2008**, *26*, 195–197. [CrossRef]
76. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
77. Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. *Adv. Intell. Comput.* **2005**, *3644*, 878–887.
78. Parsa, A.B.; Movahedi, A.; Taghipour, H.; Derrible, S.; Mohammadian, A. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accid. Anal. Prev.* **2019**, *136*, 105405. [CrossRef] [PubMed]
79. Visani, G.; Bagli, E.; Chesani, F.; Poluzzi, A.; Capuzzo, D. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *J. Oper. Res. Soc.* **2020**, *73*, 91–101. [CrossRef]
80. Hatwell, J.; Gaber, M.M.; Azad, R.M.A. Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 250. [CrossRef]
81. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
82. Dhanabal, S.; Chandramathi, S. A review of various K-nearest neighbor query processing techniques. *Int. J. Comput. Appl. Technol.* **2011**, *31*, 14–22.