Supervised and Ensemble Learning Models for Predictive Analysis (A CASE STUDY OF COVID-19)

Aishat Ojikutu[1]

Department of computer science, School of Computing and Engineering, University of Huddersfield
Huddersfield, UK

Keywords Coronavirus disease (COVID-19), Machine learning (ML), Supervised, Unsupervised Learning, Ensemble Learning, Artificial Intelligence.

I. INTRODUCTION

The novel coronavirus SARS-CoV-2 causes COVID-19, a highly infectious respiratory ailment. The illness was discovered in December 2019 in Wuhan, China and has since spread to become a global epidemic, infecting millions of people worldwide. [1]. When an infected individual talks, coughs, or sneezes, the virus spreads mostly through respiratory droplets. COVID-19 symptoms can range from moderate to severe and include fever, cough, lethargy, and trouble breathing. Some people may be asymptomatic, which means they have no symptoms of the condition. [7].

Health authorities have recommended several preventive measures, such as wearing masks, maintaining social distance, and frequent hand washing, to minimize the spread of the virus. In addition, vaccines have been developed and authorized for emergency use in many countries, which have demonstrated significant effectiveness in reducing the severity of the disease, as well as preventing hospitalization and death. [7]. There are extremely few COVID-19 test kits accessible in hospitals, which are insufficient for the rising number of patients. As a result, an autonomous detection system is needed to prevent COVID-19 from spreading among individuals.

In the fight against the COVID-19, Artificial intelligence (AI) is the primary instrument. AI contains subgroups such as machine learning (ML) and deep learning, and they all have multiple uses in various niches such as diagnosis and forecasts in the medical area owing to their accuracy. COVID-19 may be predicted and diagnosed using deep learning and (ML) approaches. There are various (ML) algorithms that can be used for these tasks, including logistic regression, decision trees, random forests, support vector machines, neural networks, and Bayesian networks. [1,2,3,7]

Logistic regression is often used to predict binary outcomes, such as whether a patient will test positive or negative for COVID-19. Decision trees and random forests are used to classify patients into distinct categories based on their symptoms and other factors. Support vector machines are used to separate patients into diverse groups based on their characteristics. Neural networks can be used to analyze complex datasets and predict outcomes such as the need for hospitalization or mortality. Bayesian networks are used to model the relationships between different variables and can be used to predict the likelihood of a patient having COVID-19 based on their symptoms and medical history.[3]. Machine learning algorithms have been shown to benefit healthcare workers in the accurate detection and prediction of COVID-19 infections, including the identification of infected patients and the monitoring of their body temperature. The RT-PCR (reverse transcription polymerase chain reaction) test is an extensively used machine method for detecting COVID-19. Other techniques include IRRCNN Model with NABLA-N, Covid-Net, and Chex-Net.

II. LITERATURE REVIEW OF PAST WORK

Several studies have explored the use of (ML) algorithms for COVID-19 outcomes prediction. The researchers in [1] used (ML) techniques [4,5,6] to understand more about how COVID-19 affects people. It was discovered that those aged 20-30, 30-40, and 40-50 are more likely to get infected with the virus. [2] aimed to provide a comprehensive review of recent research work conducted to predict the spread of COVID-19 using (ML) algorithms. The authors highlighted the different

approaches used for predicting the spread of the disease, such as mathematical models, time-series analysis, and (ML) algorithms. In line with their work [2,7] used supervised and unsupervised learning methods to predict the number of confirmed cases and deaths due to COVID-19. The findings in [2] revealed that supervised learning provides a more accurate result for recognizing COVID-19 instances, and majority of the studies used supervised learning which were about 92.1% and unsupervised learning method was about 7.1%. The study in [7] developed supervised ML models using decision tree, logistic regression, and naive Bayes algorithms to detect COVID-19 infections in Mexico. The models were trained on 80% of labeled data and tested on the remaining 20%, with the decision tree model achieving the highest accuracy of 94.99%. However, the SVM and naive Bayes models performed best in terms of sensitivity and specificity, with rates of 93.34% and 94.30%, respectively, among all models developed. In [13], a research study was conducted to determine the most effective (ML) methods for disease prediction and detection. In terms of accuracy, the study discovered that supervised learning algorithms outperformed alternative algorithms. The primary objective of the research was to demonstrate that these predictive models would be useful in combatting the COVID-19 pandemic in the future. Similarly, [3] used random forests, decision trees, and artificial neural networks to predict the number of confirmed cases and deaths due to COVID-19. The authors concluded that (ML) algorithms can accurately predict the outbreak of COVID-19 if the appropriate features are used for prediction.

The (ML) technique Extreme Gradient Boosting (XGBoost) was used in [10] to predict in-hospital mortality and critical events for COVID-19 patients at different time intervals. The system was developed using EHR data from five New York City hospitals, totaling 4098 patients. With an AUC-ROC of 0.89 for mortality prediction at 3 days and 0.80 for critical event prediction at 3 days, XGBoost beat baseline models. The strongest predictors of critical events were acute renal damage, elevated LDH, tachypnea, and hyperglycemia, whereas the strongest predictors of mortality were older age, anion gap, and C-reactive protein. Externally and prospectively, the model was also verified. This study gives encouraging evidence for the use of XGBoost in predicting COVID-19 patient outcomes. In [14] the XGBoost (ML) technique was also used to develop an accurate predictive model, it explored various hyperparameters and data processing techniques to identify the optimal model, and subsequently utilized Shapley values to determine the most informative predictors of the diagnosis. The study evaluated the model using anonymized patient data, collected from standard RT-PCR tests and additional laboratory tests, and found that the best model exhibited high diagnostic performance.

[11] presents a comprehensive survey of current research on using mobile IoT devices, AI, and telemedicine to detect and predict COVID-19. It covers various monitoring and detection methods, contact tracing, machine learning-based approaches, telemedicine for diagnosis, and privacy protection. [11] emphasizes the need to use all available tools to combat COVID-19 and highlights the challenges that lie ahead. Furthermore, it outlines future work that can be done to improve the effectiveness of using mobile IoT devices in the fight against COVID-19.

In [12] data was collected from over 11,000 COVID-19 patients admitted to Northwell Health hospitals between March and May 2020. The researchers used the data to develop and test three predictive models for 48-hour respiratory failure, comparing them to an established early warning score. The XGBoost model had the highest accuracy, outperforming the other models and the early warning score. Important predictors included oxygen delivery method, patient age, severity index level, respiratory rate, serum lactate, and demographics. The XGBoost model is a promising tool for predicting respiratory failure in COVID-19 patients.

III. SUMMARY OF THE REVIEWED LITERATURE

According to [8,7], the choice of the appropriate (ML) algorithm and the features used for prediction significantly impact the prediction accuracy. The authors suggested that the quality of input data can be improved by collecting data from multiple sources, such as health organizations, social media, and news reports. It is important to note that each algorithm has its strengths and weaknesses and may be more suitable for certain types of data or problems. Therefore, it is important to choose the right algorithm for the specific task at hand and to properly validate the results. The downside of [13] is that it requires a lot of time, specialized laboratory requirement and sometimes limitation by the unavailability of reagents required. Based on the research carried out it has been reported that new AI models with incredible forecast and detection features need to be developed in a bid to address the new and emerging strains of COVID-19. Many researchers have made significant efforts to develop machine learning models for diagnosing and predicting COVID-19. However, a critical review of the literature has revealed several methodological flaws and biases, resulting in over-optimistic reported performance. Based on current evidence, none of the machine learning models analyzed are suitable for clinical use. To increase the likelihood of developing models that can be integrated into clinical trials and validated for accuracy and cost-effectiveness, it is crucial to use high-quality datasets, adequately documented manuscripts, and external validation. Additionally, (ML) algorithms should not be used in isolation and should be combined with other methods, such as traditional statistical methods and expert knowledge, to ensure accurate predictions. However, ML approaches face a few problems, such as the newly accessible bad database. For example, One of the difficulties in training a model or selecting the best (ML) model for prediction is determining the suitable parameters.. The best (ML) model that fits the dataset was used by the researchers to make predictions based on the supplied dataset. [2].

These methods have shown promise in accurately identifying infected individuals and predicting disease progression. However, further research is needed to refine and validate these models for effective use in clinical settings.

IV. METHODOLOGY.

This investigation aims to evaluate the use of (ML) for COVID-19 diagnosis using an open-source dataset of anonymized patient data from the Hospital Israelita Albert Einstein in São Paulo, Brazil. The dataset includes samples collected during emergency room visits for SARS-CoV-2 RT-PCR and laboratory tests. The data underwent hill-climbing with various sampling approaches and hyper-parameter optimization. All clinical data were standardized to a mean of zero and a unit standard deviation, and anonymized following international guidelines.

V. DATA PREPROCESSING.

The dataset contains information about various medical parameters and laboratory test results for a group of anonymized patients. The dataset is made of 5644 entries and 111 features which includes data from blood and urine tests. The blood tests include standard measures like red blood cell count, platelet count, white blood cell count, and lymphocyte count, as well as data on the presence of different viruses such as Influenza A and B, Rhinovirus, and coronaviruses.
The urine tests make up a smaller proportion of the features, but include important measures like urobilinogen, ketone bodies, and esterase. However, some of these urine measures had inconsistencies or missing values, which necessitated the removal of some columns.
To ensure the accuracy of the data, some features had to be excluded from analysis. certain columns had an extremely high percentage of missing or invalid data, making them unsuitable for further analysis. In such cases, the columns were removed to ensure data quality. Also rows with greater than or equals to 0.80 threshold of missing values were also removed. Not done was replaced with Nan, Similarly, irrelevant features such as patients ID, the features that were not important to the predictive analysis such as patients admitted to regular ward, patient admitted into semi-intensive unit and patient admitted into the intensive care unit were also excluded from the dataset. Also the target variable was highly imbalance, for this reason I removed from the negative use cases which contains multiple null values greater than or equal to 70%. Duplicated columns were dropped, and the rest of the Nan values were filled using the KNN imputer for the MCR and MAR categories on the scikit library. Additionally, some of the columns were originally in Portuguese and had to be translated into English using google translator and the values were converted to numerical values accordingly Some of the Boolean features were represented using different string values such as True or False or 0 and 1, which were converted to a native Boolean representation. String and object-based features were also converted to integers to normalize the labels so that they only contain values between 0 and 1 These measures were taken to ensure that the dataset is accurate and consistent, and to ensure that it can be used effectively in machine learning and data analysis.
A new CSV file was created after the computation was completed and the file was used for further Analysis.

VI. DATA BALANCING.

From the chart in Fig. 1, it is obvious that the dataset is imbalance with approximately 90% of the observations tested negative for SARS-Cov-2. The remaining approximately 10% tested positive; The chart in Fig. 1, shows how this discrepancy becomes clear and visible. In practice, this interferes with the performance of the model, therefore it will be important to balance the data before feeding the final classification model. After removing the negative target that greater than or equal to 70% null values the class became 20% positive and 75% negative as shown in Fig.3. The data was further balanced using SMOTE (Synthetic minority over-sampling technique) it is a technique that uses machine learning to address class imbalance by generating synthetic samples of the minority class to balance the class distribution, This can help machine learning models perform better.
In Python, SMOTE can be implemented using the imbalanced-learn library, which provides a SMOTE class. The SMOTE class takes several parameters, such as sampling strategy to specify the ratio of the number of samples in the minority class to the majority class after resampling, and K neighbors to specify the number of nearest neighbors to use when generating synthetic samples. A new CSV file was created and saved for further Analysis after the data has been balanced. The chart in Fig. 2 shows the visualization of the target variable after the data has been balanced using SMOTE.

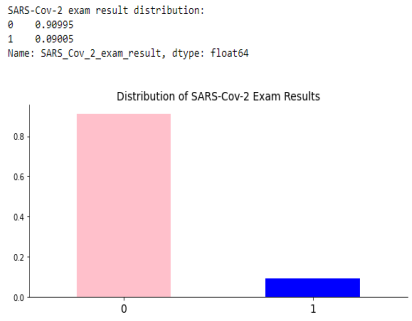


Fig. 1

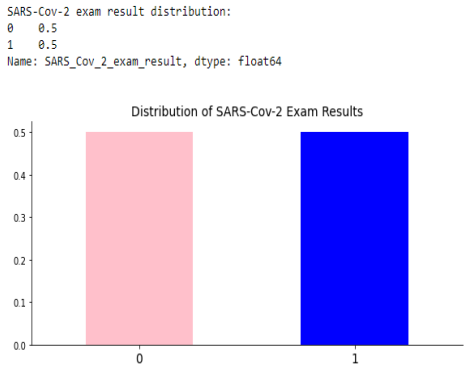


Fig. 2

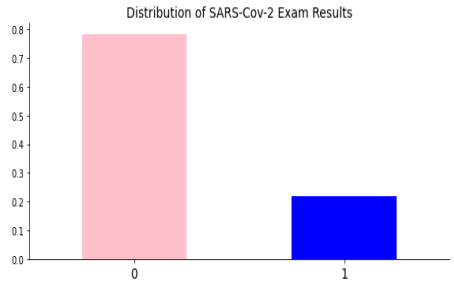


Fig. 3

VII. DATA SPLITTING

Data splitting is the practice of dividing a dataset into two or more separate subsets: a training set and a testing set in machine learning. Data splitting is used to evaluate a machine learning model's performance on new, previously unseen data. The model is trained using the training set, and its performance is evaluated using the testing set. By doing so, we can estimate how well the model will perform on new, unknown data.

The data was divided into training and testing sets using an 80/20 split. This means that 80% of the data is used for training and just 20% is used for testing.

The key point in data splitting is to ensure that the testing set is truly representative of new, unseen data. The target variable was removed in the first train and test set while the latter has just the target variable in its train and set, the two are put together into the train test split and the classifier can then be selected.

VIII. MODEL SELECTION

The prediction model was developed using Supervised Learning Model (Decision Tree classifier, Support Vector classification (SVC)) and Ensemble Learning Model (Random Forest Classifier, XG Boost Classifier and AdaBoost Classifier)

In model selection, cross-validation, information criteria, and performance metrics such as accuracy, precision, recall, AUC-ROC curve, confusion Matrix and F1-score were considered.

IX. MODEL TRAINING AND EVALUATION

To evaluate the model, the split dataset (training and testing) was fitted into the model to train the data and make predictions on the testing data. In the evaluation the model generated Accuracy, confusion matrix which exhibits TP, FP, FN, TN, the specificity which is TN/ (TN+ FP), the sensitivity which is the recall_score (t_test, y_pred), the mean and standard deviation was also gotten using the K-fold cross validation. This process was carried out for the Decision tree, XGBoost and AdaBoost Classifiers.

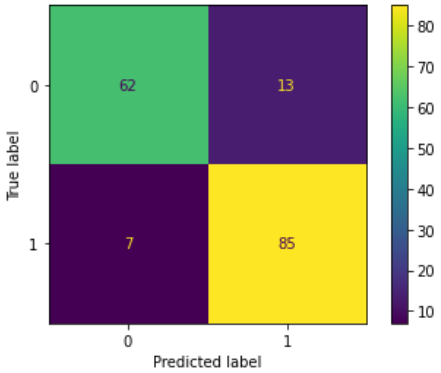
The ROC AUC, the Accuracy, and confusion matrix were also derived in the model evaluation of the SVC and Random Forest classifiers.

Below are the formulas used to calculate the performance measures utilized in this study as described in [17-24].

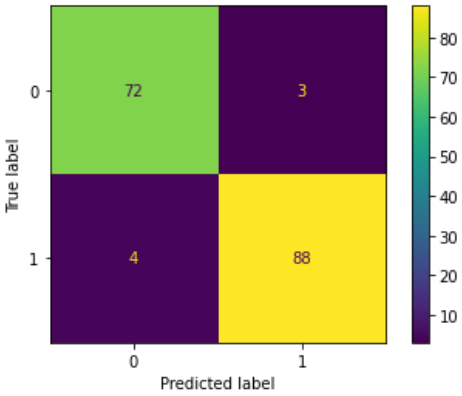
True Positive (TP) refers to cases that are truly positive and are also classified as positive by the model. False Positive (FP) refers to cases that are negative but are classified as positive. False Negative (FN) refers to cases that are positive but are classified as negative by the model. Lastly, True Negative (TN) refers to cases that are truly negative and are also classified as negative by the model.

Precision = tp/(tp+fp)
Recall = tp/(tp+fn)
F-measure = (precision × recall × 2)/(precision+recall)
Accuracy = (tp+tn)/(tp+tn+fp+fn)

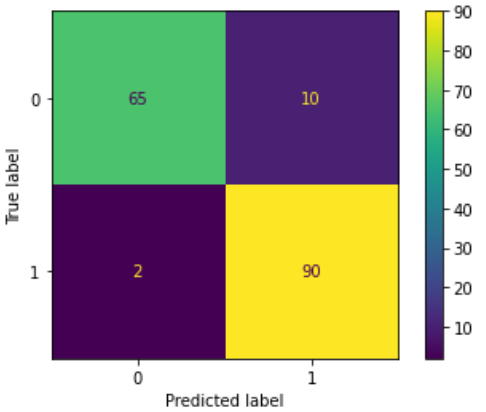
The F1 score for each model was calculated and the plots for the confusion matrix and ROC curve were displayed using the plot function which is displayed in the figures below.



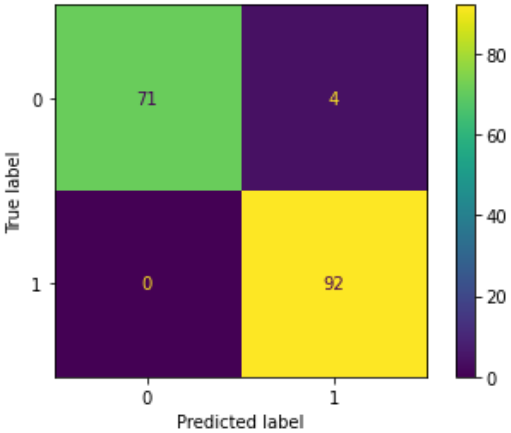
Decision Tree Classifier confusion matrix plot



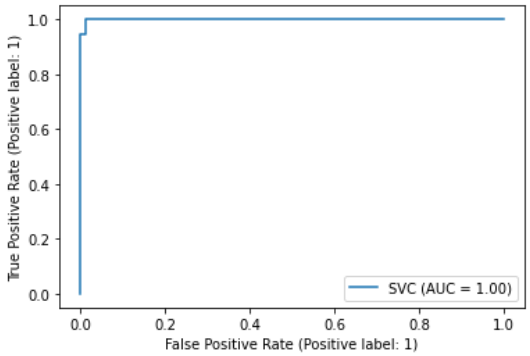
XGBoost Classifier confusion matrix plot



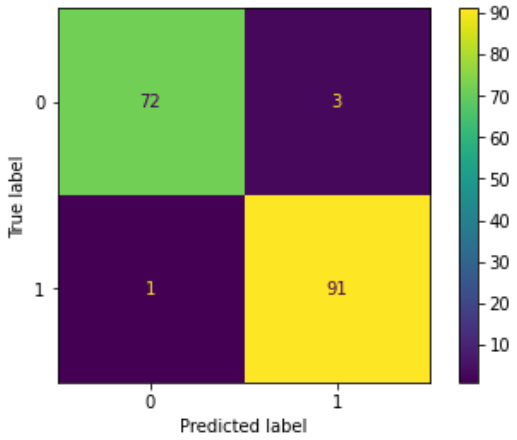
AdaBoost Classifier Confusion Matrix plot



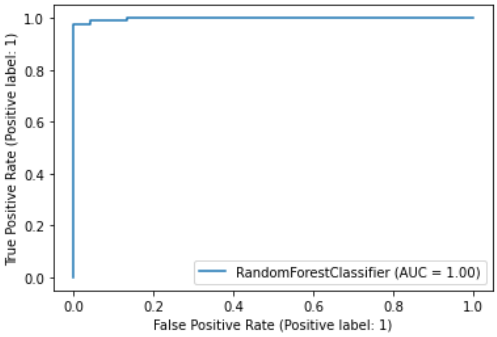
SVC Confusion matrix Plot



SVC ROC AUC curve



Random Forest confusion matrix plot



Random Forest ROC AUC Curve

The Decision Tree classifier was also used to obtain the top twenty-five key features doctors can look out for when predicting COVID-19 cases. Fig. 4 shows the visualization of the top twenty-five key features using decision tree classifier.

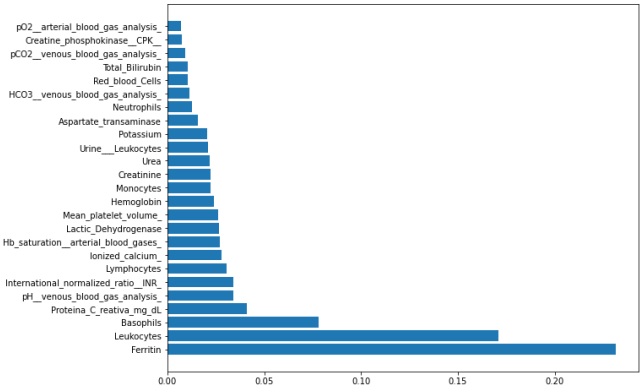


Fig.4.

By looking at the Leukocytes, Ferritin, basophils, and the rest in the order that they follow it is almost certain that the feature helps to discriminate the data indeed. The feature can bring insights for doctors when analyzing a patient. But based on past research and [1] age quantile and blood test are supposed to have a high class, these can be classified as one of the future works that needs to be done before these models can be qualified for medical diagnosis use.

Models	Accuracy %	F1 SCORE	Confusion matrix (167) [TN FP FN TP]
DT	88.02	0.91	[63 13 7 85]
XGB	95.81	0.96	[72 3 4 88]
AdaBoost	92.81	0.94	[65 10 2 90]
SVC	98.0	0.98	[71 4 0 92]
Random Forest	98.0	0.98	[72 3 1 91]

Table. 1.
Table 1 shows the accuracy, F1 score and confusion matrix of each classifier.

X. CONCLUSION

Overall, the classification model worked well, but there is definitely potential for improvement in terms of decreasing false positives and negatives. The model earned a high F1 score of 0.98 and an accuracy of 98.0%, which is highly promising. In terms of accuracy, the Support Vector Machine (SVM) and Random Forest models beat the other classifiers. In terms of F1 scores, the SVM and Random Forest models outperformed the others as well. The confusion matrix demonstrated that SVM produced the best results, with no false negatives, which is critical in COVID-19 diagnosis. This work intends to broaden by building a new algorithm based on existing models for early illness identification, specifically for COVID-19. Furthermore, data quality may be enhanced by gathering data from a variety of sources, such as health organisations, social media, different ethnicities, and geographical zones, so that the improved work is not biassed..

XI. REFERENCES

[1] Volume 8. No. 5, May 2020 International Journal of Emerging Trends in Engineering Research Available Online at https://doi.org/10.30534/ijeter/2020/117852020

[2] Kwekha-Rashid, A.S., Abduljabbar, H.N. & Alhayani, B. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. *Appl Nanosci* **13**, 2013–2025 (2023). https://doi.org/10.1007/s13204-021-01868-7 https://doi.org/10.54489/ijtim.v1i2.22

[3] A. . Akhtar, S. . Akhtar, B. . Bakhtawar, A. A. . Kashif, N. . Aziz, and M. S. . Javeid, "COVID-19 Detection from CBC using Machine Learning Techniques ", *Int. J. TIM*, vol. 1, no. 2, pp. 65–78, Dec. 2021.

[4] Balika J. Chelliah, S. Kalaiarasi, Apoorva Anand, Janakiram G, Bhaghi Rathi, Nakul K. Warrier, Classification of Mushrooms using Supervised Learning Models, IJETER, Vol 6(4), April 2018, ISSN 2454-6410, pp 229-232.

[5] D. Shona, A.Shobana, Fast and Effective Network Intrusion Detection Technique Using Hybrid Revised Algorithms, IJETER, Vol 4(11), November 2016, ISSN 2454-6410, pp 42-46.

[6] Detecting Central Nervous System Disorder Using Machine Learning Technique (XGB Classifier), Sri Lasya Dharmapuri, Pavan Kumar Dandamudi, Vinoothna Manohar Botcha and Bhanu Prakash Kolla, IJETER, Vol 8(4), April 2020, ISSN 2454-6410, pp 1142-47.

[7] Muhammad, L.J., Algehyne, E.A., Usman, S.S. et al. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN COMPUT. SCI. 2, 11 (2021). https://doi.org/10.1007/s42979-020-00394-7

[8] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing. 2006;70(1):489–501

[9] A. . Akhtar, S. . Akhtar, B. . Bakhtawar, A. A. . Kashif, N. . Aziz, and M. S. . Javeid, "COVID-19 Detection from CBC using Machine Learning Techniques ", *Int. J. TIM*, vol. 1, no. 2, pp. 65–78, Dec. 2021.

[10] Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, Johnson KW, Lee SJ, Miotto R, Richter F, Zhao S, Beckmann ND, Naik N, Kia A, Timsina P, Lala A, Paranjpe M, Golden E, Danieletto M, Singh M, Meyer D, O'Reilly PF, Huckins L, Kovatch P, Finkelstein J, Freeman RM, Argulian E, Kasarskis A, Percha B, Aberg JA, Bagiella E, Horowitz CR, Murphy B, Nestler EJ, Schadt EE, Cho JH, Cordon-Cardo C, Fuster V, Charney DS, Reich DL, Bottinger EP, Levin MA, Narula J, Fayad ZA, Just AC, Charney AW, Nadkarni GN, Glicksberg BS. Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. J Med Internet Res. 2020 Nov 6;22(11):e24018. doi: 10.2196/24018. PMID: 33027032; PMCID: PMC7652593.

[11] Shen J, Ghatti S, Levkov NR, Shen H, Sen T, Rheuban K, Enfield K, Facteau NR, Engel G, Dowdell K. A survey of COVID-19 detection and prediction approaches using mobile devices, AI, and telemedicine. Front Artif Intell. 2022 Dec 2;5:1034732. doi: 10.3389/frai.2022.1034732. PMID: 36530356; PMCID: PMC9755752.

[12] Bolourani S, Brenner M, Wang P, McGinn T, Hirsch J, Barnaby D, Zanos T, Northwell COVID-19 Research Consortium A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation J Med Internet Res 2021;23(2):e24246 https://www.jmir.org/2021/2/e24246 DOI:10.2196/24246

[13] Abirami, R.S., Kumar, G.S. Comparative Study Based on Analysis of Coronavirus Disease (COVID-19) Detection and Prediction Using Machine Learning Models. *SN COMPUT. SCI.* **3**, 79 (2022). https://doi.org/10.1007/s42979-021-00965-2

[14] Dinacci, M., Chen, T., Mahmud, M., Parkinson, S. (2022). A Case Study of Using Machine Learning Techniques for COVID-19 Diagnosis. In: Chen, T., Carter, J., Mahmud, M., Khuman, A.S. (eds) Artificial Intelligence in Healthcare. Brain Informatics and Health. Springer, Singapore. https://doi.org/10.1007/978-981-19-5272-2_10

[15] Li, D., Wang, D., Dong, J., Wang, N., Huang, H., Xu, H., & Xia, C. (2020). False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: role of deep-learning-based CT diagnosis and insights from two cases. Korean journal of radiology, 21(4), 505-508.

[16] Burog, A. I. L. D., Yacapin, C. P. R. C., Maglente, R. R. O., Macalalad-Josue, A. A., Uy, E. J. B., Dans, A. L., & Dans, L. F. (2020). Should IgM/IgG rapid test kit be used in the diagnosis of COVID19. Asia Pacific Center for Evidence Based Healthcare, 4, 1-12.

[17] [21] Iqbal, A., & Aftab, S. (2019). A Feed-Forward and Pattern Recognition ANN Model for Network Intrusion Detection. International Journal of Computer Network & Information Security, 11(4).

[18] Iqbal, A., Aftab, S., Ullah, I., Saeed, M. A., & Husen, A. (2019). A Classification Framework to Detect DoS Attacks. International Journal of Computer Network & Information Security, 11(9).

[19] Iqbal, A., Aftab, S., Ali, U., Nawaz, Z., Sana, L., Ahmad, M., & Husen, A. (2019). Performance analysis of machine learning techniques on software defect prediction using NASA datasets. Int. J. Adv. Comput. Sci. Appl, 10(5), 300-308.

[20] Iqbal, A., Aftab, S., Ullah, I., Bashir, M. S., & Saeed, M. A. (2019). A feature selection based ensemble classification framework for software defect prediction. International Journal of Modern Education and Computer Science, 11(9), 54.

[21] Iqbal, A., Aftab, S., & Matloob, F. (2019). Performance analysis of resampling techniques on class imbalance issue in software defect prediction. Int. J. Inf. Technol. Comput. Sci, 11(11), 44-53.

[22] Matloob, F., Aftab, S., & Iqbal, A. (2019). A Framework for Software Defect Prediction Using Feature Selection and Ensemble Learning Techniques. International Journal of Modern Education & Computer Science, 11(12).

[23] Iqbal, A., & Aftab, S. (2020). A Classification Framework for Software Defect Prediction Using Multi-filter Feature Selection Technique and MLP. International Journal of Modern Education & Computer Science, 12(1).

[24] Iqbal, A., & Aftab, S. (2020). Prediction of Defect Prone Software Modules using MLP based Ensemble Techniques. International Journal of Information Technology and Computer Science