Using Machine Learning for Identifying COVID-19

Shaheena Noor[1]*, Aneeta Siddiqui[2], Imran Saleem[3], Humera Noor Minhas[4]

Abstract: **In late December 2019, an outbreak of the novel coronavirus, known as COVID-19, originated in Wuhan, China, ultimately evolving into a global pandemic. This study focuses on the application of two distinct machine learning approaches to predict COVID-19 presence in individuals. The dataset employed for analysis was obtained from clients who sought medical attention at Israelita Albert Einstein Hospital in São Paulo, Brazil. During their hospital visits, samples were collected for COVID-19 and additional laboratory tests. Specifically, we utilized supervised learning techniques, namely multilayer perceptron and random forest, to conduct our analysis. The performance of these models was evaluated using the Area under the ROC Curve (AUC), resulting in respective AUC values of 97% and 92.4% for multilayer perceptron and random forest methods.**

***KeyWords*— Machine Learning; Multilayer Perceptron; Random Forest; COVID-19; Supervised Learning**

I. INTRODUCTION

COVID-19 has required remarkable reactions and extensions from the whole world and has affected millions of people out there. Governments all over the world have taken major steps like isolating the infected citizens, and imposing lockdowns to have control on the spread of the disease.

However, due to the COVID-19 symptomatology [1], the major issue is to enforce restrictions of isolation and to separate positive and negative cases of COVID-19.

For the recognition of positive cases, test is important to distinguish the infected people and hence curb the pandemic. Early diagnosis of SARS-CoV-2 is of fundamental significance to regulate the unfold of the infection. In any case, customary SARS-CoV-2 detection addicted to RT-PCR examines are often exorbitant and generally inaccessible. Despite the fact that chest CTs are found highly co-related with COVID-19 [2], this kind of test is scarcely used for screening tasks, because of the high radiation dosages, and high operation prices.

Another attempt was made using chest x-beams [3], which are low-dose and more cost-effective test. This approach had more promising measurable execution (e.g., affectability 97%). However, frameworks dependent on this test have to be altogether approved in real-world settings [4][5]. For this situation, it's hard to identify the cases throughout the incubation timeframe and they can possibly turn out to be super spreaders [6].

Although ideal confinement measures will adequately fell the unfold of the infection, the first identification will empower patients to urge additional dynamic treatment at the underlying section of the presentation [7][8]. The haematological and biochemical parameters have an impact on the underlying screening for COVID-19 [9]. In this work, we use machine learning to handle these vital parameters for exact and speedy identification of the rising sickness.

The paper follows this organization: Section 2 provides a literature overview of machine learning techniques previously utilized for COVID-19 detection. In Section 3, we present our methodology, elaborating on the machine learning algorithms employed. Section 4 covers the experiments conducted, the dataset utilized, and the obtained results. Lastly, Section 5 presents the conclusion drawn from our research.

II. LITERATURE REVIEW

2.1 Machine Learning in Medical Diagnosis

In [10], Ebru Aydndag Bayrak et al. employ Machine Learning techniques to diagnose and classify Breast Cancer using the Wisconsin Breast Cancer dataset. They compare two distinct approaches: Support Vector Machine, with SMO (Sequential Minimal Optimization) and LibSVM, as well as Artificial Neural Network, utilizing Multi-Layered Perceptron (MLP) and Voted Perceptron and achieve an accuracy of 96.99%. They use the WEKA machine learning tool for simulation.

Wu Chieh-Chen et al. in [11] have used different machine learning algorithms for the prediction of initial fatty liver screening. ROC is used to evaluate the performance of used classifiers. Among the 377 patients with liver disease, a dataset of 577 patients were included during this study. The ROC with 10-fold cross validation of RF, ANN, NB & LR

Sir Syed University of Engineering and Technology [1-2-3] | Machine Learning Engineering Group, Eyeo GmbH, Cologne, Germany[4]
Country : Pakistan , Germany
Email: shanoor@ssuet.edu.pk

was found to be 0.95, 0.925, 0.854 & 0.8888. The comparison has been created among the results and found that RF performed better than the other models.

The result forecast of patients can incredibly assist with customizing disease treatment. A lot of quantitative features (e.g., clinical tests and imaging) are conceivably helpful to evaluate the results. Picking the foremost prescient set of features is the biggest challenge. Desbordes Paul et al. in [12] have proposed GARF i.e. The study utilized a Genetic Algorithm based on Random Forest (GARF) as a feature selection technique to extract relevant features from Positron Emission Tomography (PET) clinical data and images. These selected characteristics were used to predict the therapeutic response and visualize the patient's endurance three years after the completion of treatment. The research focused on a population of sixty-five patients with locally advanced esophageal cancer eligible for chemo-radiation. The most relevant prognostic outcomes were obtained using eight selected features. Each discerning and prognostic outcomes show most well-liked performance victimisation GARF over utilizing four alternative contemplated techniques.

In [13] Yang Lei et al. suggest a way of shaping patient-specific pseudo-CT (pCT) from regularly collected MRI obsessed with data focused primarily on random forest and self-setting refinement. In order to provide and enhance semantic information, anatomical features are integrated into classification forest. A series of regression forests addicted to the auto-setting model are then accustomed to the connection of semantic information with anatomical features. The newly arrived MRI PCT is developed by separating anatomical characteristics and feeding them into the PCT prediction system. The planned algorithmic software was analysed using data set with brain scans of eleven patients.

The current artificial intelligence revolution plays a crucial role in transforming clinical practice by offering new possibilities in the field of medical images. The extensive research on machine learning for medical images indicates the significance of this area. However, it is essential to recognize that mere growth in research does not automatically lead to clinical advancements. Addressing challenges related to data requires thoughtful consideration of various factors, including the selection of datasets for projects and studies. For instance, some studies focused on classifying ADHD based on brain imaging have encountered issues like circular analysis, where feature selection was performed on the entire dataset.

Choosing appropriate datasets for research and challenges, as well as employing effective cross-validation techniques, becomes paramount in tackling these problems. Moreover, when it comes to deploying machine learning in clinical practice [14], it is vital to be cautious about relying solely on empirical results, as they may be subject to chance. Surprisingly, despite the vast number of algorithms used in clinical settings, only a meagre 6% of them are truly effective health interventions.
A revealing observation from a study involving 410 highly downloaded papers from the ACM is that approximately 97% of the papers utilizing significance testing reported findings with p-values lower than 0.0562. This highlights the importance of interpreting statistical results carefully and avoiding overreliance on marginal p-values to draw significant conclusions. By being mindful of these aspects, we can ensure that machine learning in medical diagnosis truly makes a meaningful impact on clinical practice.

Ghulab Nabi Ahmad et.al [15] focuses on the challenging task of predicting cardiac disease, which demands significant time and effort from medical professionals. The study delves into the implementation of diverse Machine Learning algorithms to achieve this prediction. To validate their system, a 5-fold cross-validation method is employed. Furthermore, the researchers conduct comparative analyses on four datasets from UCI Kaggle. This evaluation allows them to assess the performance of various methodologies utilized in their study.

The findings reveal that the Extreme Gradient Boosting classifier, in combination with GridSearchCV, achieves the highest accuracies in both testing and training, scoring 100% and 99.03% respectively. Moreover, even without GridSearchCV, the XGBoost Classifier demonstrates remarkable performance. In essence, the main objective of this paper is to present an innovative approach for creating models that effectively tackle real-world challenges, with a specific focus on cardiac disease prediction.

Breast cancer stands as a significant global health issue affecting women, and in this article, the author introduces a novel expert system designed for diagnosing the disease [16]. The system relies on RBF-KELM comes with adjustable parameters, specifically the penalty parameter C and the parameter (σ) of the RBF-kernel., which significantly impact its efficiency. In this study, the researchers used the DE algorithm to obtain optimal values for these parameters. The obtained results showcased the approach's effectiveness, surpassing conventional methods and providing satisfactory diagnostic outcomes.

2.2 Role of Machine Learning in COVID-19 pandemic

The unexpected episode of corona virus sickness has quickly become a significant worldwide health issue. Jiangpeng Wu in [17] used random forest algorithmic rule to extract eleven blood indices to construct the ultimate assistant tool from forty-nine clinical available biopsy information. In order to precisely differentiate COVID-19 from the related assortment of suspicious patients with equivalent CT details or identical symptoms, the technique introduced robust findings with a

precision of 0.9795 and 0.9697 for the cross-validation and testing collection. The proposed device is suitable to complete primer analysis.

Davide Brinati in [18] have created 2 ML-based models for classification utilizing values from blood tests (to be specific: the platelets, white blood cells counts, LDH plasma levels, ALP, GGT, ALT, AST, CRP) of 279 patients who presented with COVID-19 symptoms were subjected to screening using the rRT-PCR test conducted on respiratory samples. lot examples. 177 of the case were positive, while the rest got a negative reaction. The accuracy ranges somewhere in the range of 82% and 86%, and affectability between 92% to 95%, so comparably well as for the best quality level. In addition, the author built an explanatory decision Tree model as a clear alternative guide for COVID-19 speculate cases for practitioners to decode blood tests (even off-line). This study highlighted the efficacy and suitability of exploring blood samples and employing machine learning as an alternative to RT-PCR for distinguishing COVID-19 patients. This is particularly valuable in countries, including developing ones, where there may be shortages of RT-PCR reagents and limited access to specialized analysis centers. The author also developed an online-based tool, accessible for clinical reference and analysis purposes.

Tao Ai et al. in [2] used Chest CT within the identification of corona virus sickness. The examination enclosed 1014 patients for Wuhan, China, which intimates between January 6th and February 6th 2020, with every chest CT and RT-PCR test. The presentation of chest CT inside the COVID-19 finding was assessed with the use of RT-PCR because of the reference norm. Similarly, the dynamic transformation of RT-PCR findings was cleft compared to patients with multiple RT-PCR measurements and successive chest CT exams for those with a four-day or additional interval between the RT-PCR analysis. In the 413 patients with negative RT-PCR results, 75% had positive chest CT discoveries. Among the 1014 patients, approximately 60% to 93% exhibited initial positive CT scans that were consistent with COVID-19 before or at the same time as the initial positive RT-PCR results.

In this investigation, Apostolopoulos, Ioannis D. et al. [3], utilize a dataset of X-ray footage from patients with basic microorganism respiratory illness, thoroughbred COVID-19 illness, and standard occurrences, was used for the automated recognition of the Corona virus ill. Convolutional Neural Networks have been used in this study to gauge the performance as it was proposed in the recent years for classification of medical images. The authors collected X-ray images from publicly available repositories of X-ray pictures to create their dataset. The findings indicate that utilizing Deep Learning with X-ray imaging can effectively resulting in impressive rates of specificity, sensitivity, and accuracy, which are 96.46%, 98.66%, and 96.78%, respectively.

The impact of Long COVID, which refers to the lingering effects of SARS-CoV-2 infection on patients, has been significant, affecting both patient recovery and society's response to the COVID-19 pandemic. In this study [19], the researchers developed machine learning models using electronic health records from the National COVID Cohort Collaborative (N3C) to identify potential long COVID patients. The focus of the study was on non-deceased adult patients who had received a COVID-19 diagnosis or tested positive for SARS-CoV-2. This approach facilitates the identification of patients who may benefit from specialized care for long COVID and helps in pinpointing suitable candidates for clinical trials, thereby contributing to a better understanding of long COVID as its definition evolves. Moreover, the models can be retrained and adjusted using data from diverse sources for future studies.

Since at this point, every single symptomatic test show disappointment rates, for example, to raise issues, the chance of incorporating X-rays into the conclusion of the unwellness can be surveyed by the clinical network, in view of the discoveries, whereas a lot of exploration to assess the X-ray approach from numerous viewpoints may well be directed.
The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has significantly impacted the global economy and healthcare systems. However, the conventional diagnostic RT-PCR method often produces false-negative and inaccurate results. In article [20], the author explores the potential of utilizing blood tests and machine learning for COVID-19 prediction. Through the application of various classifiers and the Synthetic Minority Oversampling Technique (SMOTE), the study aims to enhance classification accuracy. The Shapley Additive Explanations (SHAP) method identifies eosinophils, monocytes, leukocytes, and platelets as crucial blood parameters for effectively distinguishing COVID-19 infection. By combining these classifiers with RT-PCR tests, sensitivity can be improved, particularly during emergencies like pandemics involving new virus strains. The promising results suggest the feasibility of implementing an automated framework to support clinicians in diagnosing and screening patients more efficiently.

COVID-19, caused by the SARS-CoV-2 virus, has become a global pandemic impacting millions of individuals worldwide. With conventional diagnostic methods facing challenges in coping with the surge in infections, researchers have turned to intelligent techniques to achieve rapid and accurate COVID-19 diagnosis. In the context of this paper [21], a comprehensive review is presented, encompassing more than 200 studies published from various publishers, exploring the latest DL and ML approaches employed in COVID-19 diagnosis. The research tracks are categorized into DL and ML, and the paper assesses COVID-19 public datasets from different countries.

Various evaluation measures, including accuracy, sensitivity, and specificity, are comparatively analysed and discussed in these studies. SVM emerges as the most widely used ML mechanism for COVID-19 diagnosis and outbreak prediction, while CNN remains the primary DL mechanism. The overarching objective of this review paper is to provide guidance to the research community, urging further advancement in ML and DL techniques for COVID-19 diagnosis and inspiring progress and innovation in this critical domain.

III. METHOD OF ANALYSIS

3.1 Training of a Multi-Layer Perceptron (MLP) for COVID-19 classification.

Machine Learning is a technology that has seen its use and popularity rise exponentially in the last few years. Based on the advantages of MLP compared with other machine learning algorithms, MLP is used to classify whether a particular patient has COVID-19 or not. MLP are inspired by Neural Network, i.e., human brains and they follow the high-level biological counterpart's structure. Normally, the training of MLP is done using the basic learning rules, which is then combined with back propagation algorithm. Consider Fig. 1, shows a 3-layer MLP, with one layer each of input, hidden and output neurons. In Algorithm1, an overview is given of our MLP algorithm for training and testing.

Algorithm 1 Training and Testing using MLP.

1: Data:
2: x_i:<B, T, P>
3: where B: Blood Parameters: {b1, b2,……..bn}
4: T:Treatment Intensity: {t1, t2,……..tn}
5: P: Personal Details: {p1, p2,……..pn}
6: C ∈{0,1}
7: Result
8: Training:
Given xi and yi;
Calculate w:
Set of weights to generate prediction;
Model M.
9. Testing
Given Model M,
Predict C i.e. presence or absence of COVID-19 in patient.
10: Training Phase:
11: Step 0: Initialize weights and bias:
W-[W1….Wi,b1……bk]
12: Repeat steps 1 and 2 for all input data points:
13: Step 1: Feedforward pass:
compute x̃i-f(Wixi+ bi)
14: Step 2: Backpropagation:
15: Compute gradients: Δ_ω J_a (W).
16: Compute weight change: ΔW .
17: Update weight W.
18: Testing Phase:
19: Given unseen datapoints {b_i,t_(i,) p_i} as input.
20: Use Model M to query the network (MLP)
21: Predict ci i.e. whether person has COVID-19 or not

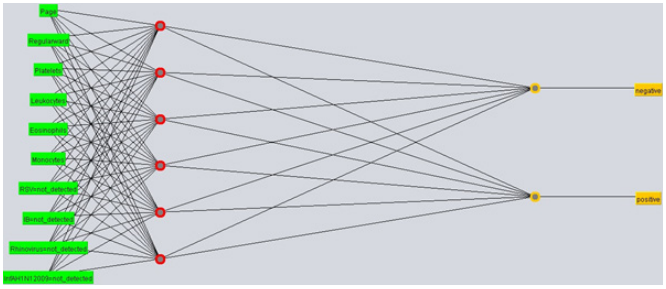

Fig. 1 Multi-layer Perceptron for detection of COVID-19

3.2 Random Forest

We have used Random Forest classifier algorithm [22] in our research which creates multiple decision trees for classification process. The algorithm divided our data into smaller subsets, at the same time by adding branches to the tree. The resultant tree contains a root node and decision nodes. The root node of each individual tree in the forest includes a bootstrap section from the original dataset as the training set. The remaining dataset which is not included in training dataset is called out-of-bag (OOB) dataset. The decision node is further divided into two or more branches which represents the value of each feature (like blood parameters, treatment intensity and personal details) tested and the leaf node has the result on the patient's prospective condition (outcome value). Random forest provides the final result by averaging the result provided by multiple decision trees.

Here h1 (x), h2 (x)…hk (x) is the group of decision trees and the training data that is obtained from the vectors X, Y;
where, X denotes the input parameters, and they are as follow:

$$X :< B, T, P >$$

B: Blood Parameters: $\{b_1, b_2, ...b_m\}$

T: Treatment Parameters: $\{t_1, t_2, ...t_n\}$

P: Personal Details: $\{p_1, p_2, ...p_o\}$

& Y denotes the output parameter $Y \in \{0,1\}$. Let's examine Eq. (1), which presents the random forest margin function. This function quantifies the extent to which the total number of votes in X, Y for the correct class surpasses the average vote for any other class. A larger margin indicates higher confidence in the classification. In simpler terms, we can interpret it as a measure of the difference between data points that are classified correctly and those that are misclassified.

$$mg(X,Y) = av_k I(h_k(x) = Y) - \max_{j \neq k} av_k I(hk(X) = j) \quad (1)$$

Where, I is the indicator function.

If $mg(X,Y) > 0$, means set of classifiers vote for the correct classification and

If $mg(X,Y) < 0$, means set of classifier vote for the incorrect classification.

The generalization error is simply the misclassification rate and is expressed in equation 2.

$$PE' = P_{X,Y}(mg(,Y) < 0) \quad (2)$$

Where x,y, indicates the space of likelihood.

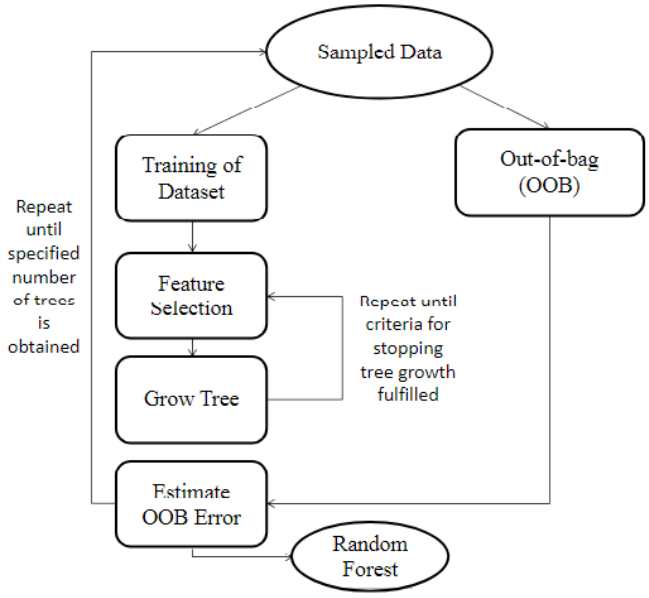

Figure 2: Random Forest

IV. EXPERIMENTS AND RESULTS

We address how we perform our experiments and the findings obtained in this section. We begin by explaining the dataset and then talk about research strategy.

4.1 Dataset

The dataset contains anonymized data from patients seen in São Paulo, Brazil at Israelita Albert Einstein Hospital who had samples of SARS-CoV-2 along with RT-PCR. In addition, more tests were done during the visit. [22]. There are 109 parameters in the dataset along with an ID and the output with prediction. The dataset has 5644 datapoints with 10% positive examples.

We have divided our dataset into three different blocks given in Figure 3 that show in detail which features or predictors we have taken as input parameters for the classification. Each block indicates the blood cell for the classification of COVID-19 and more details of the condition of a person and intensity of treatment received. First of all, we did data pre-processing to clean up the dataset. As mentioned above, the total input parameters in the original dataset are 109, however, we noticed that (1) all the parameters are not (equally) important for classification (2) there are a lot of missing values.

Hence, we divided the dataset into 3 blocks to find out the important parameters and clean-up accordingly. The partitioning of the complete dataset is driven by following principles:

1. Removed parameters from too many missing data points (>= 95%).
2. Samples that are too sparse are often omitted from laboratory records. Negative samples have been retained that have at least 10 variables with available data points. This is done to prevent an over-fit situation where there might be an excessive effect on the predictive model of a few samples (sparse but positive).

Accordingly, Block 1 comprises of 18 features. Block 2 has 35 features against the original dataset in [22], where the author uses 44 predictors. Finally, we have
Block 3 comprising of 10 features, which are the same as defined by the author in [22].

Block	Features considered
Block 1	Patient age quantile, Patient admitted to regular ward, Patient admitted to semi-intensive unit, Patient admitted to intensive care unit, Hematocrit, Hemoglobin, Platelets, Mean platelet volume, Red blood Cells, Lymphocytes, Mean corpuscular hemoglobin concentration (MCHC), Leukocytes, Basophils, Mean corpuscular hemoglobin (MCH), Eosinophils, Mean corpuscular volume (MCV), Monocytes, Red blood cell distribution width (RDW).
Block 2	Patient age quantile, Patient admitted to regular ward, Patient admitted to semi-intensive unit, Patient admitted to intensive care unit, Hematocrit, Hemoglobin, Platelets, Mean platelet volume, Red blood Cells, Lymphocytes, Mean corpuscular hemoglobin concentration (MCHC), Leukocytes, Basophils, Mean corpuscular hemoglobin (MCH), Eosinophils, Mean corpuscular volume (MCV), Monocytes, Red blood cell distribution width (RDW), Respiratory Syncytial Virus Influenza A, Influenza B, Parainfluenza 1, CoronavirusNL63, Rhinovirus/ Enterovirus, Coronavirus HKU1, Parainfluenza 3, Chlamdophila pneumonia, Adenovirus, Parainfluenza 4, Coronavirus229E, CoronavirusOC43, Inf A H1N1 2009, Bordetella pertussis, Metapneumovirus, Parainfluenza 2. NOTE: In addition to above parameters, the author Tharsis Souza et al. also consider the following parameters. Neutrophils, Urea, Proteina Creativa mg dL, Creatinine, Potassium, Sodium, Influenza.B rapid.test, Influenza.A rapid.test, Strepto.A
Block 3	Patient age quantile, Patient admitted to regular ward, Platelets, Leukocytes, Eosinophils, Monocytes, Respiratory Syncytial Virus, Influenza B, Rhinovirus, InfA H1N12009

4.2 Training and Results

Consider Table 1, where we show the configuration of MLP used to train our network. We are using 3 blocks of dataset with a different set of input parameters; therefore, the number of input neurons are different for each case. For example, Block1, Block 2 and Block 3 has 18, 35 and 10 input neurons respectively. We are using 1 hidden layer with different number of neurons based on input; while we are using 2 output neurons that classify the presence or absence of COVID-19.

The training is done using sigmoid transfer function. With a rate of learning of 0.3 and a momentum of 0.2, it took 0.69s, 1.24s and 0.18s to fully train the network for each respective block. For Random Forest, a single thread is used to ensemble a batch size of 100 with a seed value of 1. The bag size is set to encompass the entire training set, which represents 100% of the data, with the number of iterations (trees) fixed at 100. Additionally, each tree is grown to an infinite depth.

The results of the COVID-19 classification are given in Table 2 and a comparison is given diagrammatically in Fig. 4. In all these 3 block configurations, we can clearly see that the precision of Block 2 is higher than Block 1 & 3 in both MLP and random forests.

Furthermore, we compared our results given in Table 3 to build the credibility of our work directly with original work [22]. The author in [22] achieves 92% AUC; while both our strategies gave better performance. Using MLP, we got an AUC of 97% and for Random Forest, we got 92.4%.

Table 1 - Details of MLP

	Block 1	Block 2	Block3
Input parameters/ neuron	18	34	10
No. of Hidden Layers (No. (of neurons	(10)1	(18)1	(6)1
Output parameters/ neurons	2	2	2
Time taken to build model ((in seconds	0.69	1.24	0.18
Momentum	0.2		
Learning Rate	0.3		

Table 2: COVID-19 Results

Machine Learning Techniques	**Block 1**			**Block 2**			**Block 3**		
	Precision	Recall	AUC	Precision	Recall	AUC	Precision	Recall	AUC
Multi Layer Perceptron	86	87.1	82.2	93.5	93.8	97	90.7	91.1	89.1
Random Forest	90.4	89.2	86.2	91.9	91	92.4	88.9	89.7	83.5

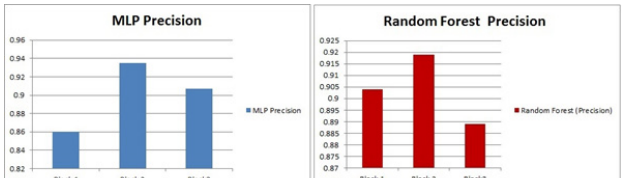

Figure 4: Precision of MLP (L) and Random Forest (R)

Table 3: Comparison of our result with Tharsis Souza [22]

Tharsis Souza et al. Result	Our Result	
AUC (%)	AUC (%)	
	MLP	Random Forest
92	97	92.4

V. CONCLUSIONS

The novel coronavirus disease, also known as COVID-19, first emerged in Wuhan, China, in late December 2019, eventually escalating into a global pandemic. Identifying the presence of the virus in individuals is a fundamental task critical to its effective treatment and containment. This study focuses on employing two machine learning approaches to predict COVID-19 infection in individuals. For this purpose, we utilized a dataset obtained from Israelita Albert Einstein Hospital in São Paulo, Brazil, where samples were collected during hospital visits specifically for COVID-19 testing. Employing supervised learning techniques, namely multilayer perceptron anandom forest algorithms, we achieved significant Area Under the Curve (AUC) values of 97% and 92.4%, respectively, in our predictive models. These results hold promise in bolstering COVID-19 detection efforts and advancing strategies to combat its spread. Further research on larger and diverse datasets can enhance the practicality and generalizability of these predictive models in real-world healthcare settings.

REFERENCES

[2] Day, M. (2020). Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village. BMJ: British Medical Journal (Online), 368.

[2] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., ... & Xia, L. (2020). Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology, 200642.

[3] Apostolopoulos, I. D., & Mpesiana, T. A. (2020). Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, 1.

[4] Weinstock, M. B., Echenique, A. N. A., DABR, J. W. R., Leib, A., & ILLUZZI, F. A. (2020). Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. J Urgent Care Med, 14(7), 13-8.

[5] Cabitza, F., & Zeitoun, J. D. (2019). The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Annals of translational medicine, 7(8).

[6] Lai, C. C., Liu, Y. H., Wang, C. Y., Wang, Y. H., Hsueh, S. C., Yen, M. Y., ... & Hsueh, P. R. (2020). Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (SARSCoV-2): facts and myths. Journal of Microbiology, Immunology and Infection.

[7] Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., ... & Zhao, Y. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. Jama, 323(11), 1061-1069.

[8] Gao, Y., Li, T., Han, M., Li, X., Wu, D., Xu, Y., ... & Wang, L. (2020). Diagnostic utility of clinical laboratory data determinations for patients with the severe COVID-19. Journal of medical virology.

[9] Liu, Y., Yang, Y., Zhang, C., Huang, F., Wang, F., Yuan, J., ... & Zhang, Z. (2020). Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury. Science China Life Sciences, 63(3), 364-374.

[10] Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (pp. 1-3). IEEE.

[11] Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. Computer methods and programs in biomedicine, 170, 23-29.

[12] Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G. (2017). Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Computerized Medical Imaging and Graphics, 60, 42-49.

[13] Lei, Y., Wang, T., Harms, J., Shafai-Erfani, G., Tian, S., Higgins, K., ... & Liu, T. (2019, March). MRI-based pseudo CT generation using classification and regression random forest. In Medical Imaging 2019: Physics of Medical Imaging (Vol. 10948, p. 1094843). International Society for Optics and Photonics.

[14] Varoquaux, Gaël, and Veronika Cheplygina. "Machine learning for medical imaging: methodological failures and recommendations for the future." NPJ digital medicine 5.1 (2022): 48.

[15] Ahmad, Ghulab Nabi, et al. "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV." IEEE Access 10 (2022): 80151-80173.

[16] Bacha, Sawssen, and Okba Taouali. "A novel machine learning approach for breast cancer diagnosis." Measurement 187 (2022): 110233.

[17] Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., ... & Zhao, M. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv.

[18] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: a Feasibility Study. medRxiv.

[19] Pfaff, Emily R., et al. "Identifying who has long COVID in the USA: a machine learning approach using N3C data." The Lancet Digital Health 4.7 (2022): e532-e541.

[20] Chadaga, Krishnaraj, et al. "Clinical and laboratory approach to diagnose COVID-19 using machine learning." Interdisciplinary sciences: Computational life sciences 14.2 (2022): 452-470.

[21] Alyasseri, Zaid Abdi Alkareem, et al. "Review on COVID-19 diagnosis models based on machine learning and deep learning approaches." Expert systems 39.3 (2022): e12759.

[22] Souza T, Gustavo WS and Souza HSP, (2020). COVID-19 Machine Learning-Based Rapid Diagnosis From Common Laboratory Tests, Online Published at: https://www.kaggle.com/felixthecat/covid-19-ml-policy-based-rapiddiagnosis