

# Evaluating Three Machine Learning Classification Methods for Effective COVID-19 Diagnosis

Akbal Omran Salman<sup>1</sup>, Oana Geman<sup>2,\*</sup>

<sup>1</sup>Department of Control & Automation Techniques Engineering, Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq; [akbal.o.salman@mtu.edu.iq](mailto:akbal.o.salman@mtu.edu.iq)

<sup>2</sup>Department of Computers, Electronics and Automation, Faculty of Electrical Engineering and Computer Science, Stefan cel Mare University of Suceava, 720229 Suceava, Romania; [oana.geman@usm.ro](mailto:oana.geman@usm.ro)

*Received 05.12.2022, Revised 26.12.2022, Accepted 09.01.2023, Published 15.01.2023*

**ABSTRACT:** SARS-CoV2, which produces COVID-19, has spread worldwide. Since the number of patients is rising daily, it requires time to evaluate laboratory data, limiting treatment and discoveries. Such restrictions necessitate a clinical decision-making tool with predictive algorithms. Predictive algorithms help healthcare systems by spotting disorders. This study uses machine learning and laboratory data to predict COVID-19 patients. Recall, Precision, accuracy, and AUC ratings assessed our models' prediction performance. Models were verified with 10-fold cross-validation and train-test split methods using 18 laboratory data from 600 patients. This research compared three different classification approaches—Support Vector Machines (SVM), artificial neural networks (ANN), and k-Nearest Neighbors (k-NN). According to the findings, SVM achieved the most significant average accuracy (89.3%), followed by ANN (88.5%) and kNN (86.6%). The accuracy rates of all three approaches were relatively reasonable, with SVM being the best of the bunch. The results of this research indicate that classification using machine learning methods has the potential to be used in developing reliable COVID-19 diagnosis systems, thereby facilitating the fast and accurate diagnosis of COVID-19 cases and facilitating proper therapy and management of COVID-19 patients. More work might be done to refine these techniques and include them in useable diagnostic frameworks.

**Keywords:** COVID-19, Machine Learning, Artificial neural network, Support Vector Machine, K-Nearest Neighbors, labtory data.

## 1. INTRODUCTION

There has been a lot of pressure on hospitals and medical centers worldwide to provide accurate diagnosis and treatment for the millions of people who have contracted COVID-19. Timely and precise detection of COVID-19 infections is a significant problem in controlling the spread of the virus. To effectively treat and manage patients, it is crucial to building accurate COVID-19 diagnostic methods. Since machine learning classification algorithms can reliably evaluate big datasets and give insights into complex trends that are difficult for human specialists to identify, they show promise in developing such systems [1].

The discipline of diagnosis has significantly benefited from AI and ML use during the COVID-19 epidemic. The necessity for precise and prompt diagnosis is a significant obstacle to preventing the spread of the virus. It can take a long time and a lot of money for conventional diagnostic procedures like PCR testing to produce findings. With the help of AI and machine learning, better and faster testing methods for COVID-19 have been created. Among these is the evaluation of X-ray and CT scan pictures of the chest for the presence of COVID-19 using artificial intelligence algorithms. Potential instances can be identified using these instruments, lessening the strain on healthcare workers. The possibility that a patient has COVID-19 may also be predicted using machine learning algorithms that assess variations in symptoms and risk variables. Such models can be invaluable for patient triage in environments with limited medical resources. The creation of quick antigen testing is another area where artificial intelligence and machine learning were utilized for COVID-19 diagnosis. These tests are helpful for mass screening and testing since artificial intelligence algorithms analyze the data in a couple of minutes [2].

When the number of people needing treatment in a hospital exceeds the number of doctors, nurses, and beds available, overcrowding becomes a severe issue. No socioeconomic group is immune to the adverse effects of this worldwide public health crisis, which includes longer wait times, worse service quality, and less efficient healthcare professionals [3]. Patient admissions and discharges, wait times, resource availability, and clustering contribute to overcrowding. It significantly contributes to rising death tolls and the maintenance of discriminatory queueing practices. The COVID-19 pandemic is worsening a bad situation by raising mortality rates and making diagnosis more difficult[3]. Due to the overlap in symptoms with other disorders, testing for COVID-19 is recommended for diagnosis. Depending on the severity of the infection, symptoms can occur anywhere from 2-14 days after contact and include fever, cough, exhaustion, shortness of breath, chest

discomfort, muscular pains, headache, loss of taste or smell, sore throat, congestion, diarrhea, nausea, vomiting, stomach pain, and rash. A lack of appetite in children may accompany these symptoms. However, delays in diagnosis might occur because the diagnostic equipment is not always reliable [3]. Therefore, effective and precise diagnostic tools are required to lessen crowding and stop the development of COVID-19. Solutions can be found in artificial intelligence and machine learning, which can facilitate faster and more accurate diagnoses, forecast patient outcomes, and maximize hospital efficiency. For instance, to ease the workload of healthcare providers, AI systems may examine X-ray and CT images for signs of COVID-19. Early treatments and the spread of the disease can be prevented thanks to machine learning's ability to forecast the possibility of a patient having COVID-19 based on symptoms and risk factors. Healthcare practitioners may save more lives and deliver better treatment in less time using artificial intelligence and machine learning.

Methods for detecting COVID-19 include:

- Serologic testing [4].
- Testing for the virus in a throat swab [5].
- Quantitative real-time reverse transcription-polymerase chain reaction.

However, earlier research has shown that chest radiography (X-rays) and chest computed tomography (CT) scans are effective at detecting anomalies indicating lung illness, particularly COVID-19 [6, 7]. X-ray and CT scans are the primary detection methods for COVID-19. They may also be used to assess the severity of the disease, monitor the emergency of infected individuals, and anticipate the course of the disease [8]. However, traditional manual diagnostics cannot be employed in such emergency scenarios due to time constraints [9]. Due to the potential for human error in the evaluation, learning, and understanding of the results, the services of a medical professional should be sought out. Hospitals worldwide are overrun with patients with varying degrees of health as COVID-19 transmission rates soar [10]. Therefore, the patient test must be executed rapidly and efficiently to preserve as many people's lives as feasible [4]. Effective Diagnosis and severity categorization of COVID-19 can be aided by intelligent technology [6].

AI is becoming increasingly popular in various contexts, especially in medical diagnosis and illness detection [11]. There has been widespread use of AI in multiple settings because it speeds up the generation of reliable detection findings while lightening the strain on healthcare infrastructure [12]. In addition, AI can shorten the time it takes to make a call compared to conventional detection methods [13]. Developing AI such that the hazards of epidemic illnesses may be recognized is a crucial strategy to enhance future global health risk identification, early detection, and diagnosis [14]. Several authors [8] have introduced several AI classifiers tested on real-world COVID-19 datasets covering a wide range of objectives and use cases. Selecting an AI approach suitable for generating correct findings remains a crucial difficulty [15, 16], notwithstanding the benefits AI algorithms have in identifying and categorizing COVID-19. Because of the abundance of available AI methods, it might be challenging to determine which one is most suited for COVID-19 diagnosis and categorization [17].

Only in research [18] have machine learning models and laboratory data been used to diagnose individuals with COVID-19. The authors made a notable contribution by balancing and filtering a dataset containing 111 test results from 5644 patients. From a sample of 600 patients, they determined that just 18 of 111 test results were clinically relevant. Multiple deep-learning models have been tried out on this dataset, with the CNN-LSTM combination achieving the highest accuracy at 92.3%. Despite these impressive results, enhanced ML models and diagnostic accuracy are still possible. Since deep learning models rely on several parameters and deep layers [19–22], deploying them in real-time applications is challenging without a significant investment in time and computing power. That is to say, and we should not expect lightweight performance from these models. As an added complication, the combined approach (CNN-LSTM) is resource-intensive and barely meets the demands of such models in real-time settings. In addition, the same study adopted a few features based on the recommendations of other studies from a medical perspective while ignoring the feature selection method depending on the specifications of ML models (technical standpoint), which is especially problematic given that the amount of COVID-19 patient data can be unpredictable and prompt medical involvement is required.

This study aims to develop an interdisciplinary COVID-19 classification approach using three robust contemporary machine learning approaches to automatically identify healthy and COVID-19-infected people based on laboratory data. The research has three main objectives:

- To investigate the role of classification algorithms in disease prediction using COVID-19.
- Three machine learning classifiers, the Support Vector Machine (SVM), artificial neural network (ANN), and the k-Nearest Neighbors (kNN) were used to classify the risk variables that influence a diagnosis of COVID-19.
- Using a dataset of COVID-19 patients and their risk variables, compare the accuracy, recall, and precision of the three machine learning classifiers.

This study's remaining sections are structured as follows. Section 2 reviews previous studies on applying machine learning classification techniques to diagnose COVID-19. Section 3 details the dataset and method of comparing the three classification approaches. The study's findings and a comparison of the three approaches are presented in Section 4; finally, the conclusion is presented in Section 5 by discussing where the research findings.

## 2. The Proposed Method

Millions of individuals have been affected by the COVID-19 epidemic, and it has caused enormous economic and social devastation throughout the world. Diagnosing COVID-19 patients quickly and correctly is a significant obstacle in stopping the epidemic. Despite developing several diagnostic procedures, such as RT-PCR and antibody assays, these tests can be laborious, costly, and insensitive. The use of machine learning methods in developing reliable COVID-19 diagnostic systems with the potential to increase diagnostic throughput and precision is an area of exciting recent research.

The SVM, KNN, and decision tree are all proposed, and their efficacy in classifying COVID-19 cases is compared in this study. This research aims to create a reliable COVID-19 diagnostic system that will aid doctors in making prompt and correct diagnoses, therefore limiting the propagation of the disease and enhancing patient outcomes. The suggested approach includes data collection and preparation for the COVID-19 patient dataset, feature extraction, COVID-19 classification, and technique evaluation. The following parts provide a more in-depth explanation of each stage of the suggested technique; the details of the proposed methodology are presented in Figure 1.

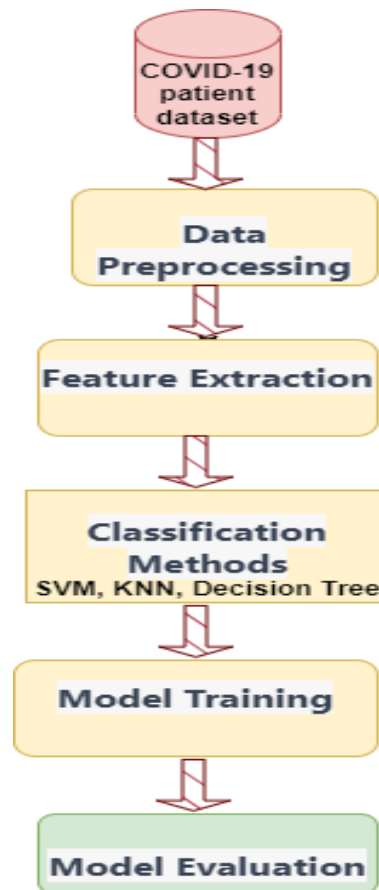


Figure 1. Evaluating three machine learning classification methods for effective COVID-19 diagnosis

### 2.1 Dataset used

The dataset, accessible via [23], contains the laboratory results of patients treated at the Hospital Israelita Albert Einstein in Sao Paulo, Brazil. Patients' samples were taken in the early 2020s for SARS-CoV2 detection. Laboratory results totaling 111 for 5644 individuals are included in the dataset. Positive patients accounted for almost 10% of the sample, with 6.5% and 2.5% requiring hospitalization and critical care, respectively. No information on gender is included in the dataset. Eighteen diagnostic tests have been shown to significantly impact COVID-19 disease [23, 24, 25]. As a result, to normalize the data and do COVID-19 detection, we removed all remaining features from the lab. Since some patients did not have access to all 18 laboratory findings, the total number of patients in the dataset was reduced from 5644 to 600 during the balancing

procedure. There are 520 patients with no results and 80 individuals with COVID-19 in the matched sample. The results of the lab work are presented in Table 1. Researchers at <https://github.com/burakalakuss/COVID-19-Clinical> may access the symmetrical dataset. The sample data of COVID-19 patients show in Figure 2.

Table 1. Eighteen results from the patients' lab tests included in the dataset

Labtory findings	Basophils transaminas, aspartate, hemoglobin, C reactive protein, serum glucose, red blood cells, potassium, sodium, platelets, monocytes, creatinine, urea, alanine transaminase, lymphocytes, hematocrit, eosinophils, neutrophils, leukocytes
------------------	--

atient age quantil	Hematocrit	Hemoglobin	Platelets	Red blood Cells	Lymphocytes	Leukocytes	Basophils
17	0.236515	-0.0223403	-0.517413	0.102004	0.318366	-0.0946103	-0.223767
1	-1.57168	-0.774212	1.42967	-0.850035	-0.00573804	0.36455	-0.223767
9	-0.747693	-0.586244	-0.42948	-1.36132	-1.11451	-0.884923	0.0816925
11	0.991838	0.792188	0.072992	0.542763	0.0454363	-0.211488	-0.834685
9	0.190738	-0.147652	-0.668155	-0.127191	0.00279103	-1.13259	0.387152
13	1.01473	0.854844	-0.178244	0.489872	-0.730707	-0.0751308	2.52537
14	0.740064	0.854844	0.361914	0.436981	-0.227493	0.105751	-0.529226
9	-0.679027	-0.711556	0.952319	-0.321124	-0.875701	-0.286623	-0.223767
8	0.236515	0.040316	0.072992	0.0667435	1.2395	-0.545423	0.99807
17	0.80873	1.04281	-0.278739	0.701437	-0.261609	4.61877e-06	0.387152
14	-0.106813	-0.33562	1.42967	-0.60321	0.326895	0.726313	0.0816925
16	1.42672	1.23078	0.600588	1.26561	0.506005	-0.0222577	0.692611
15	0.717175	0.60422	0.399599	0.40172	0.0113201	-0.178094	0.387152
17	1.01473	0.666876	-0.203368	0.295938	-0.261609	-0.0834792	-0.223767
7	0.946061	1.16812	-0.278739	1.17746	-0.671003	0.133579	0.692611
19	0.625621	0.353596	-0.127997	0.190156	-1.59214	0.556563	0.0816925
6	-0.15259	-0.460932	1.24124	-0.250603	-0.0654413	1.19661	-0.529226
10	0.190738	0.040316	-1.05757	-0.462168	-1.26804	1.22443	-0.529226
11	0.740064	0.60422	0.160925	-0.0214086	-0.329842	0.225411	-0.223767

Figure 2: samples data of COVID-19 paitents

## 2.2 Preprocessing stage

To build a reliable COVID-19 diagnosis system, it is necessary to preprocess the data using three different machine learning classification algorithms and then compare their results. Accurate results can only be achieved by using high-quality data for training and testing classification algorithms. The term "data preprocessing" refers to the procedures performed on raw data to clean, convert, and ready it for analysis.

Data cleaning is the initial stage of the preprocessing phase, and it entails the elimination of extraneous information and the completion of any gaps in the data. The information is then "normalized" so that all scales are consistent. This step is essential for machine learning models to avoid favoring variables with more significant rankings. Next, relevant features are retrieved from the preprocessed data in a process known as feature selection. Reducing the number of variables and zeroing down on the most critical predictors is the goal of this stage, as doing so can boost the efficiency of models created using machine learning. Data transformation, the process of turning data into a format more conducive to machine learning algorithms, is another crucial stage. To better train machine learning models, it might be helpful to transform categorical data into numerical data using one-hot encoding or transform text data into numbers using word embedding. A reliable COVID-19 diagnostic system relies heavily on the quality of the data that has been preprocessed. The precision of the diagnostic system may be enhanced by cleaning, converting, and testing machine learning models on high-quality data.

## 2.3 Feature Extraction stage

This effort aims to present an enhanced COVID-19 identification model utilizing novel ML techniques. No studies have combined IoT and several ML methods for predicting COVID-19 from laboratory data. The current investigation may inspire further studies to validate the approaches using more laboratory data. Using a brute-force approach, we found the most crucial laboratory characteristics to enhance accuracy and choose the optimal model. The prediction performance of ML systems is impacted by irrelevant features, making feature selection essential. Feature selection improves prediction accuracy and speeds up the ML algorithm running. The brute-force feature selection method exhaustively evaluates all possible combinations to find the

best input characteristics. Overfitting is a significant concern because of the prohibitively high computational cost of thorough searches. As a result, people resort to avaricious strategies like forward resolve out of desperation.

Comparing the variation in the accuracy of ML techniques in the COVID-19 lab results highlighted the significance of feature selection. We eliminated monocytes, salt, and alanine transaminase from the original list of 18 clinical characteristics and kept the top 15, which agrees with the medical perspective on specific characteristics but not all. The features of the dataset are presented in Table 2.

Table 2 The features of the dataset

No.	Name of Features	Types of Features
1	Patient age quantile	Integer
2	Hematocrit	Integer
3	Hemoglobin	Integer
4	Platelets	Integer
5	Red blood Cells	Integer
6	Lymphocytes	Integer
7	Leukocytes	Integer
8	Basophils	Integer
9	Eosinophils	Integer
10	Monocytes	Integer
11	Serum Glucose	Integer
12	Neutrophils	Integer
13	Urea	Integer
14	Protein C reativa mg/dl	Integer
15	Creatinine	Integer
16	Potassium	Integer
17	Sodium	Integer
18	Alanine transaminase	Integer
19	Aspartame transaminase	Integer
20	Label	Boolean

## 2.4 COVID-19 Classification

Algorithms powered by AI may analyze past data and extrapolate future results. It is possible to classify ML algorithms as a subset of AI. It's a field focused on self-improvement by studying computer algorithms for learning and optimization. Some distinctions separate deep learning from machine learning. Powerful computing and complexity were barriers for DL algorithms until recently. However, advancements in big data have enabled more profound and extensive networks, allowing computers to learn, monitor, and react to complicated events more quickly than people. Here, we create and assess clinical prediction models for correlating test evidence of COVID-19 infection with clinical diagnosis. We trained SVMs, RFs, and kNNs to compare how well each method evaluated the study's results. ANN is a method of processing data that takes cues from the human brain's organic nerve system. Neurons, activation functions, input, output, and hidden layers all make up this structure.

Data mining classification allocates a group element to a target class. Classification predicts the target category for each data set condition. In taxonomy, the inputs are divided into two or more categories, and the learner must produce a model that allocates invisible inputs to one or more. In classification, they have many algorithms or methods to determine each method's performance. Each method has a different performance to know it gets accuracy. Examples of classification methods are Decision Trees, Support Vector Machine, Decision forests, Neural Networks, Nave Baysin, Gradient Boosting Machines (Augmented Decision Tree), and more. Figure 3 shows the classification method used in this study.

They are inspired by how the human brain functions; the ANNs are a particular type of machine-learning algorithm. They have a wide variety of applications, one of which is categorizing COVID-19-related clinical data. For COVID-19 clinical data classification, ANNs are helpful because they can learn detailed mappings between feature inputs and label outputs. As a result of the variability in symptoms and test findings, doctors may need help to correctly diagnose and treat individuals with COVID-19. ANNs can sift through mountains of clinical data and spot trends and correlations humans would miss. This could increase the precision of diagnoses, the effectiveness of treatments, and the efficiency with which resources are allocated. In addition, public health professionals can keep better tabs on outbreaks and react faster because of ANNs' ability to evaluate and understand data about COVID-19 in real-time. In general, ANNs have an opportunity to

completely transform the categorization of COVID-19 clinical data and enhance the standard of care provided to patients.

Classification of COVID-19 clinical data also benefits from using Support Vector Machines (SVMs). Supervised learning algorithms like SVMs may classify data by identifying the optimal hyperplane as a dividing line between groups. SVMs may be applied to the clinical data from COVID-19 to discover associations between patient characteristics and outcomes like illness progression or mortality. As frequently in clinical situations where numerous features must be assessed concurrently, SVMs excel at handling high-dimensional data. In addition, SVMs can increase classification accuracy by taking non-linear data using a variety of kernel functions. To further aid clinical decision-making, SVMs can also give a confidence estimate for their predictions. Because of their ability to detect intricate interrelationships and patterns in massive, high-dimensional datasets, SVMs have great potential for COVID-19 clinical data categorization. SVMs can produce precise and trustworthy predictions that can enhance the results for patients and public health responses when used with additional machine learning methods, including feature selection and ensemble approaches.

When classifying COVID-19 clinical data, K-Nearest Neighbors (KNN) is another helpful technique. K-Nearest Neighbors (KNN) is a supervised machine learning technique that uses inter-point distance, for instance, labeling. KNN may be used to find commonalities between various clinical characteristics and outcomes, such as the prognosis of illness severity or death, in the setting of COVID-19 clinical data. KNN's straightforward design and simplicity are two of its main selling points. It works with numeric and categorical variables and makes no assumptions about the data's underlying distribution.

KNN may also be utilized for multi-class classification applications and has variable decision bounds. KNN also benefits from being easily interpretable. For example, the KNN method generates a list of the K nearest neighbors for a new instance to better understand how various attributes and outcomes are related. As a result, significant risk indicators for COVID-19 can be identified, which can aid in therapeutic decision-making. They are building SVM, ANN, and KNN models presented in Figures 4, 5, and 6.

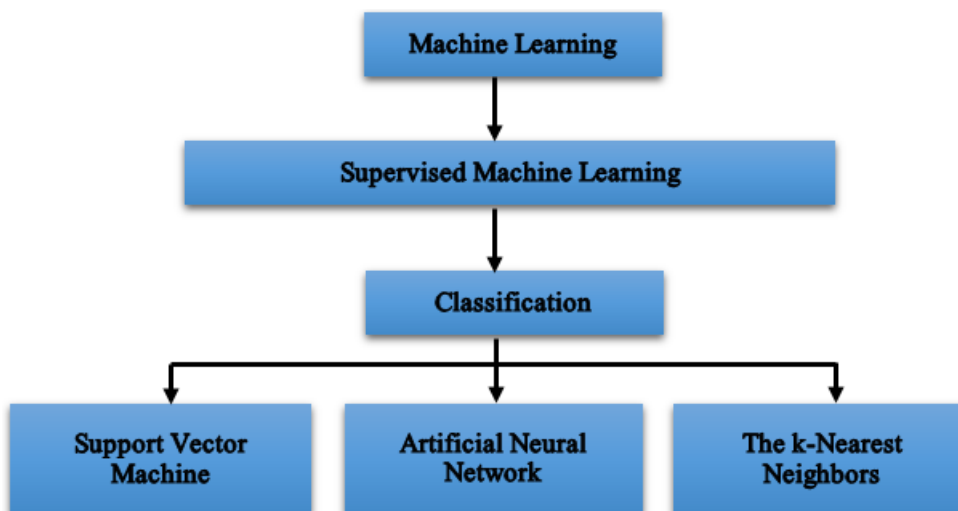


Figure 3: COVID-19 classification Methods

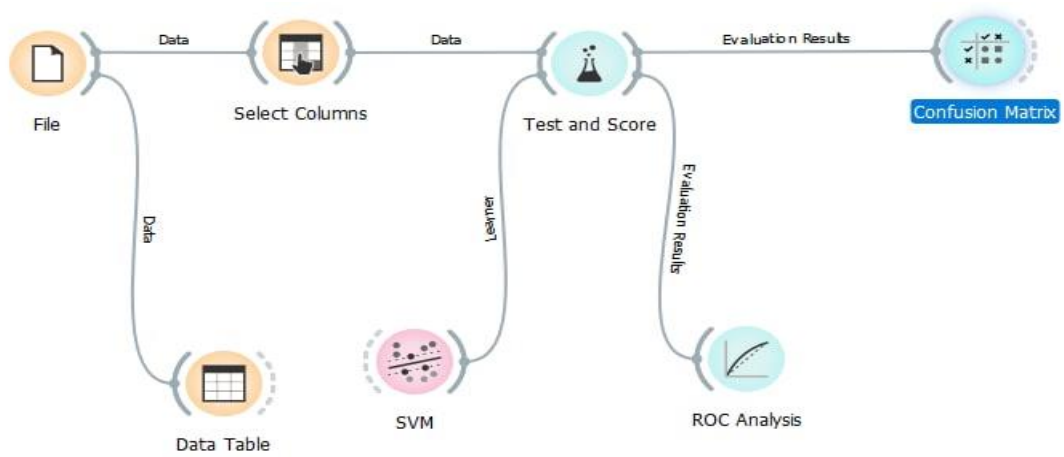


Figure 4: Building a support vector machine model

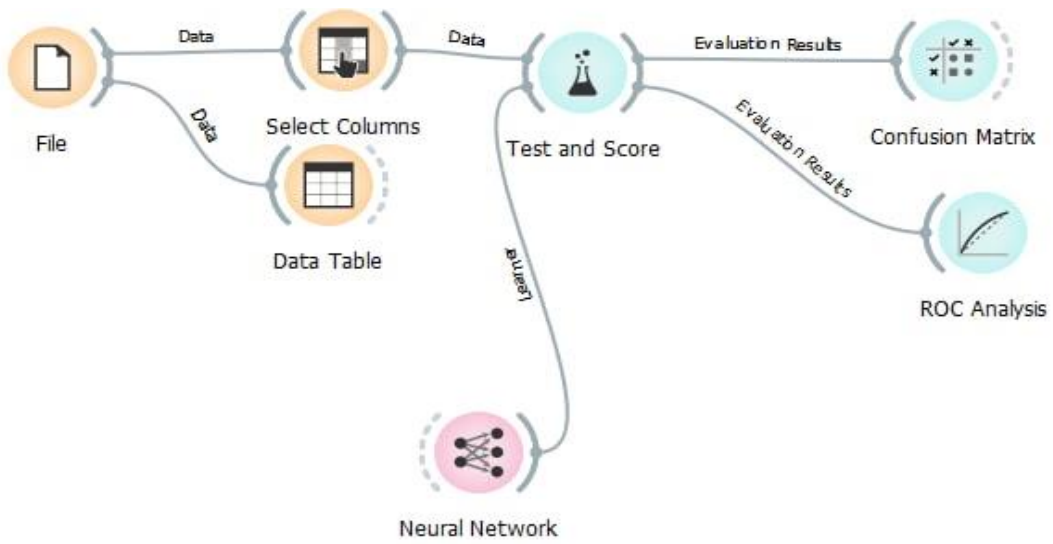


Figure 5: Building Artificial Neural Network model

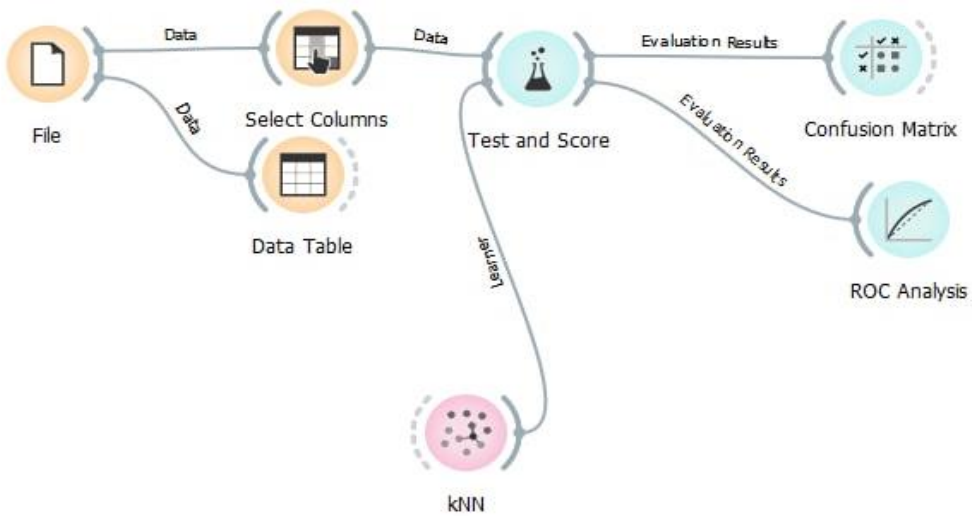


Figure 6: Building The k-Nearest Neighbors model

## 2.5 Evaluation Metric

In experiments, accuracy, precision, and recall are evaluated using a confusion matrix. When predicting  $m$  classes, the confusion matrix is  $m \times m$ . The evaluation matrix rows represent target classes, and columns represent output classes. Select an appropriate threshold to call the incident excellent or negative. Classifiers identify instances as positive or negative based on their likelihood. As indicated in Table 3, the assessment metric was generated using a confusion matrix to quantify how effectively categorization differed from the output.

Table 3: Confusion Matrix

Confusion Matrix	Class 1	Class 2
	Predicted	Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Classes 1 and 2 allow for both positive and negative solutions, hence the equations may be defined as:

- Positive (P): The result of the observation is favorable (a positive diagnosis of Covid-19, for instance).
- Negative (N): Neither a positive nor a negative observation (such as a negative diagnosis of Covid-19).
- If an observation matches what was expected, we say that it is a True Positive (TP).
- False Positive (FP): The observation matches the null hypothesis.
- Observation is negative, as expected; this is a True Negative (TN).
- False Positive (FP): When the actual result differs from the expected one.

The fraction of correctly predicted cases is defined by the first equation in the definition of precision. It is also defined as the ratio of the number of positive samples that were correctly identified to the entire number of samples.

$$\text{Precision, } P = \frac{TP}{TP + FP} \quad (1)$$

Recall or Sensitivity or True positive rate (TPR) as in equation 2 is the proportion of positive cases that were correctly identified. It also defined by the ratio of the total number of correctly classified positive samples divided by predicting the total number of positive samples.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

Accuracy as in equation 3 is the proportion of the total number of prediction that was correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

## 3. RESULTS AND DISCUSSION

Using medical imaging and other clinical data, machine learning classification methods have been proven to diagnose COVID-19 accurately. Using these techniques, clinicians may rapidly and reliably examine massive volumes of data, improving patient treatment quality. Table 4 displays the accuracy results for SVM, ANN, and The KNN. Ten experiments with varying cross-validation fold data allocations have been run to compare the performance of SVM, KNN, and ANN. According to the results, SVM performs best when split into 90% training and 10% testing data, with the maximum accuracy score of 0.892% in the first test. The accuracy drops to 0.870% in test number 10 when the data is split 33% in favor of training and 100% in favor of testing. The SVM produces high-quality classification results, with an average accuracy score of 0.8827 and a standard deviation of 0.0068.

Table 4: The Accuracy results of the SVM, KNN and ANN

Test	Split data	SVM	KNN	ANN
1	90:10	0.892	0.883	0.888
2	95:20	0.883	0.882	0.882
3	80:20	0.889	0.878	0.884
4	75:50	0.888	0.878	0.888
5	70:50	0.887	0.880	0.890



6	66:50	0.886	0.880	0.889
7	60:50	0.882	0.877	0.883
8	50:50	0.876	0.877	0.884
9	40:100	0.874	0.874	0.879
10	33:100	0.870	0.872	0.882
Average		0.8827	0.8781	0.8849
Standard Deviation		0.0068	0.0032	0.0034

The results also show that the first test, with 90% training data and 10% testing data, yielded the most excellent accuracy for KNN. There's a maximum of 0.883% accuracy. The tenth test yielded the worst results, with a data split of 33% training and 100% testing. The worst possible accuracy rating is 0.872%. The average accuracy score for SVM and ANN is 0.8781%, but KNN classification results tend to be significantly lower than those. KNN's standard deviation yields a value of 0.0032. Results for SVM, KNN, and ANN accuracy are shown in Figure 7.

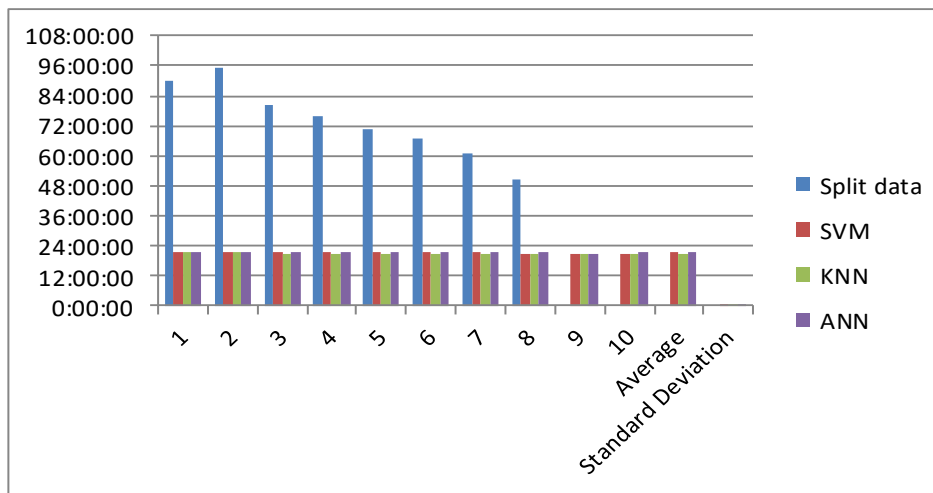


Figure 7: The Accuracy results of the SVM, KNN and ANN

Precision test results for SVM, KNN, and ANN are shown in Table 5. Ten experiments with varying cross-validation fold data allocations have been run to compare the performance of SVM, KNN, and ANN. Based on the results of the tests, the optimal accuracy for SVM is achieved when the data is split 90% for training and 10% for testing (test 1). On the other hand, the accuracy drops to 0.860% in test #9, when the data was split between 40% training and 100% testing. The average precision score for SVM classifications is 0.8732, with a standard deviation of 0.00722. This is significantly higher than the results obtained using other approaches.

Table 5: The Precision results of the SVM, KNN and ANN

Test	Split data	SVM	KNN	ANN
1	90:10	0.883	0.864	0.878
2	95:20	0.868	0.861	0.869
3	80:20	0.878	0.855	0.874
4	75:50	0.881	0.856	0.877
5	70:50	0.878	0.858	0.880
6	66:50	0.878	0.858	0.878
7	60:50	0.873	0.852	0.872
8	50:50	0.868	0.852	0.872
9	40:100	0.860	0.846	0.865
10	33:100	0.865	0.844	0.865
Average		0.8732	0.8546	0.873
Standard Deviation		0.00722	0.00595	0.00488

The results also show that KNN performs best when the data is divided into 90% training and 10% testing. Accuracy of 0.864% is the best that can be achieved. With an average KNN precision score of 0.8546 and a standard deviation of 0.00595, Test 10's data split of 67% for training and 33% for testing produces the lowest precision score of 0.844. According to the findings, ANN performs at its peak when data is divided 70% for training and 50% for testing. Up to 0.880% accuracy may be achieved. The accuracy was worse in tests nine and ten, which used data splits of 40% training and 100% training and 33% training and 100% testing, respectively. Standard deviation is 0.00488 points, with an average ANN precision of 0.873%. The lowest accuracy possible is 0.865. Figure 8: Accuracy Evaluations of SVM, KNN, and ANN.

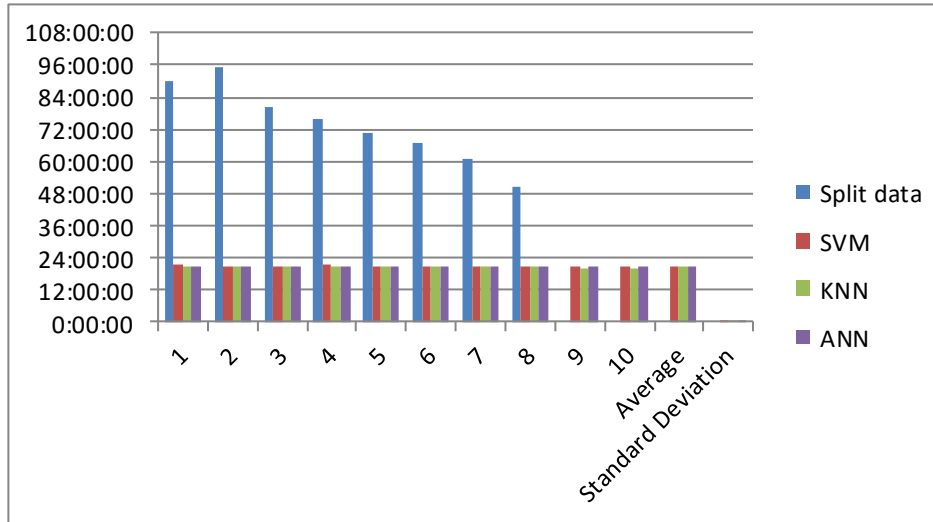


Figure 8: The Precision results of the SVM, KNN and ANN

As can be shown in Table 6, when the data is divided 90% for training and 10% for testing, Test 1 yields the greatest level of recall for SVM, with a score of 0.892. The lowest recall achieved using SVM was 0.870 on Test 10, which uses all testing data and none of the training data. With an average recall value of 0.8827 and a standard deviation of 0.00682, the SVM are usually reliable.

Table 6: The recall results of the SVM, KNN and ANN

Test	Split data	SVM	KNN	ANN
1	90:10	0.892	0.883	0.888
2	95:20	0.883	0.882	0.882
3	80:20	0.889	0.878	0.884
4	75:50	0.888	0.878	0.888
5	70:50	0.887	0.880	0.890
6	66:50	0.886	0.880	0.889
7	60:50	0.882	0.877	0.883
8	50:50	0.876	0.877	0.884
9	40:100	0.874	0.874	0.879
10	33:100	0.870	0.872	0.882
Average		0.8827	0.8781	0.8849
Standard Deviation		0.00682	0.00261	0.00344

The results also show that the optimal recall for KNN is obtained in the initial test when the data is divided into 90% training and 10% testing. Maximum recall is 0.883 percent. Because the data was divided between 33 percent training and 100 percent testing, Test 10's recall of 0.872 was the lowest of any test. KNN classification has

an average recall score of 0.8781 and a standard deviation of 0.00261. This is slightly lower than the results achieved by SVM and ANN.

According to the results, the optimal recall for ANN is achieved when the data is split 70% for training and 30% for testing. The highest possible recall is 0.890. Tests two and ten, with data split 95% training and 20% training and 33% training and 100% testing, respectively, had the weakest recall, with a score of 0.882. The average recall score for the Gradient Boosted Machine's classification results is 0.8849, while the standard deviation for the

ANN's recall is 0.00344. Figure 9: Recall Results for SVM, KNN, and ANN Algorithms.

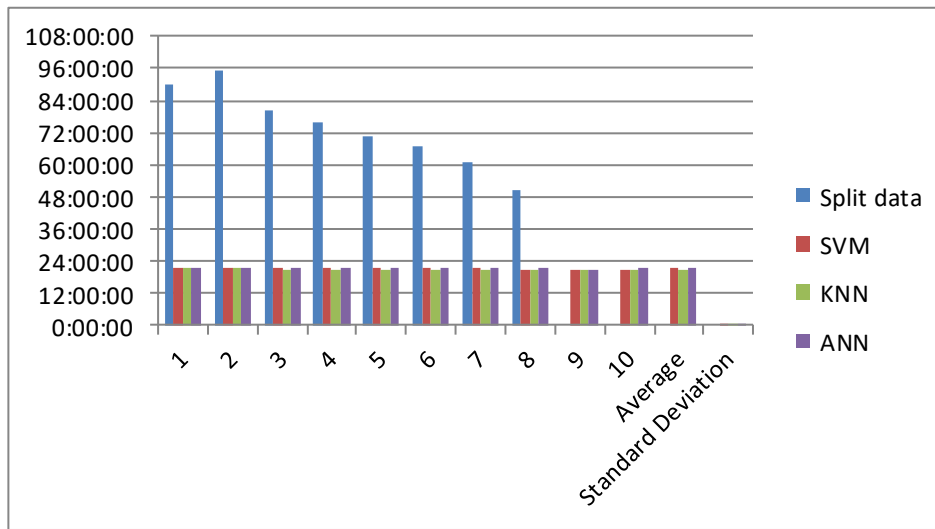


Figure 9: Result of Recall for SVM, KNN and ANN algorithms

More generally, the diagnostic analysis assesses the precision of a statistical model (e.g., logistic regression, linear discriminant analysis) that classifies an individual as either healthy or ill. Its purpose is to provide a straightforward visual representation of a diagnostic test's reliability. The outcomes of a curve analysis test are often either discrete (binary) or ordinal (ordinarily) scales. Screening tests often produce binary findings, with  $D_x = 0$  for healthy and  $D_x = 1$  for ill patients, presented in Figures 10 and 11. A patient's likelihood of having a disease is determined via a screening test. The evaluated information might be ordinal if more than two categories are used.

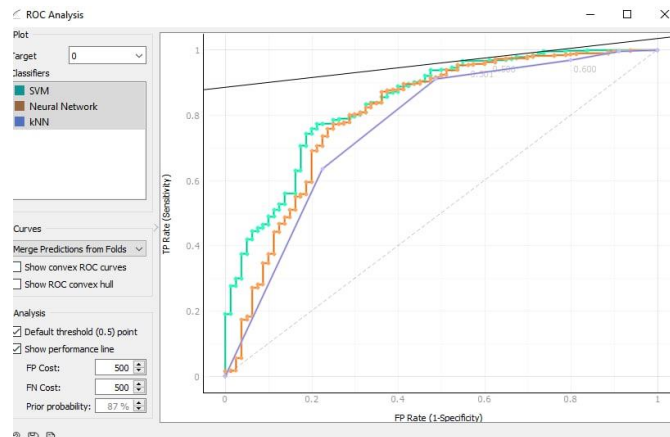


Figure 10: The test indicates an uninfected patient  $DX = 0$

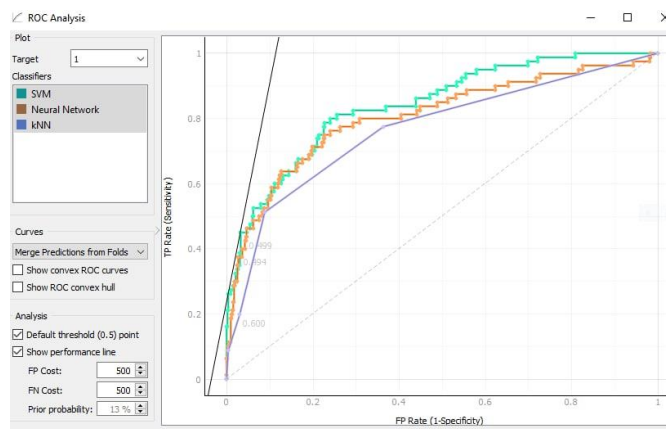


Figure 11: The test indicates the patient has  $DX=1$

In this study, the proposed algorithms' efficacy has been tested against the Coronavirus dataset's risk factor. The three approaches to this project have been selected for comparison and evaluation. These approaches are SVM, ANN, and KNN. As it is shown, Table 7 shows the comparison between the previous studies with deep learning models and the current research. Those studies use the same dataset, which is the Risk Factor of the coronavirus. Previous research used sex methods: ANN, CNN, RNN, LSTM, CNNLSTM, and CNNRNN. In comparison, the current study uses three techniques from previous research: SVM, KNN, and ANN.

The previous studies use 6-fold cross-validation to evaluate the performance of ANN, CNN, RNN, LSTM, CNNLSTM, and CNNRNN. The current research uses 10-fold cross-validation to assess the performance of SVM, ANN, and KNN. In the previous study, they separated each method with different feature numbers to know the accuracy of each method. While for the current project using the number of the same features, which are 20 features for each method, to see the accuracy. Based on Table 4.5 below, both projects have similarities where SVM method accuracy is higher than other methods.

Table 7: The comparison between previous studies with a current proposed approach

	Previous Project (Comparison of deep learning approaches to predict COVID-19 infection)	Current Project
<b>Dataset</b>	Corona virus(Risk Factors) Data Set.	Corona virus(Risk Factors) Data Set.

Method	ANN,CNN,RNN,LSTM,CNNLSTM,CNNRNN							SVM,ANN,KNN			
K-fold	10-fold validation							10-fold validation			
Accuracy		ANN	CNN	RNN	LSTM.	CNNLSTM	CNNRNN	SVM	KNN	ANN	
	Features Numbers	20	20	20	20	20	20	20	20	20	20
	Accuracy (%)	0.8600	0.8800	0.8416	0.8666	0.8416	0.8566	0.893	0.867	0.885	

The amount of data is the primary limitation of this study. Some test results could not be accurately quantified in the sample of 600 patients. However, the forecast was accurate between 82% and 94% of the time within a statistically significant population. The data also needed to be more balanced; therefore, we eliminated certain materials to increase the proportion. Having more data is critical to improving these models' efficiency.

#### 4. CONCLUSION

In this study, we used machine learning models trained on laboratory data to forecast the spread of COVID-19. As mentioned earlier, three machine learning models were used to examine the laboratory—the initial step of the research involved standardizing the data and feeding it into machine learning methods. The classification was then run, and the models' efficacies were evaluated by measures including precision, recall, accuracy, and area under the curve (AUC). We used train-test split and 10-fold cross-validation to ensure our models were accurate. The risk factor for the covid-19 data set was used for the classification experiment and was carried out using this data. This study aims to classify the risk factor for the COVID-19 dataset. The results are presented based on the performance of the algorithm and the accuracy in classifying the risk factor for the COVID-19 dataset. In this project, different machine learning methods, ANN, SVM, and kNN methods, were implemented and applied to the risk factor of the COVID-19 dataset. Hence, the classification model based on the vector machine support classifier is an effective classifier with an accuracy of 89.3% for classifying COVID-19 data. For future works, this research will employ specific well-known feature selection approaches to improve the classification algorithm's effectiveness in picking out relevant characteristics. It would be best to look at several methods of increasing performance precision. More research using laboratory data collected from other facilities is required to confirm these findings. Only samples from Israelita Albert Einstein Hospital were tested. Furthermore, the prediction effectiveness of the models may be impacted by the existence of varying illness phases. Values like fever and lymphopenia were less critical in the prediction process than other variables in this study, suggesting that decision-making systems may discriminate between patients and healthy individuals. Early detection of COVID-19 disorders and early treatment options can be offered in future research with the adoption of AI algorithms and the rise in the volume of data.

**Funding:** This research received no external funding.

**Conflict of interest:** The authors declare no conflicts of interest.

#### REFERENCES

- [1] Abdulkareem, K.H., Mohammed, M.A., Salim, A., Arif, M., Geman, O., Gupta, D. and Khanna, A., 2021. Realizing an effective COVID-19 diagnosis system based on machine learning and IOT in smart hospital environment. *IEEE Internet of things journal*, 8(21), pp.15919-15928.
- [2] Mohammed, M.A., Maashi, M.S., Arif, M., Nallapaneni, M.K. and Geman, O., 2021. Intelligent systems and computational methods in medical and healthcare solutions with their challenges during COVID-19 pandemic. *Journal of Intelligent Systems*, 30(1), pp.976-979.
- [3] Dinar, A.M., Raheem, E.A., Abdulkareem, K.H., Mohammed, M.A., Oleiwie, M.G., Zayr, F.H., Al-Boridi, O., Al-Mhiqani, M.N. and Al-Andoli, M.N., 2022. Towards Automated Multiclass Severity Prediction Approach for COVID-19 Infections Based on Combinations of Clinical Data. *Mobile Information Systems*, 2022.
- [4] H. Zeng et al., "Antibodies in infants born to mothers with COVID-19 pneumonia," vol. 323, no. 18, pp. 1848-1849, 2020.
- [5] D. A. J. A. o. p. Schwartz and I. medicine, "An analysis of 38 pregnant women with COVID-19, their newborn infants, and maternal-fetal transmission of SARS-CoV-2: maternal coronavirus infections and pregnancy outcomes," vol. 144, no. 7, pp. 799-805, 2020.
- [6] T. Ozturk et al., "Automated detection of COVID-19 cases using deep neural networks with X-ray images," p. 103792, 2020.
- [7] M. Li et al., "Coronavirus disease (COVID-19): spectrum of CT findings and temporal progression of the disease," 2020.
- [8] F. Shi et al., "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," 2020.
- [9] H. S. Maghddid, A. T. Asaad, K. Z. Ghafour, A. S. Sadiq, and M. K. J. a. p. a. Khan, "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms," 2020.
- [10] W. H. Organization, "Coronavirus disease 2019 (COVID-19): situation report, 72," 2020.
- [11] D. D. Miller and E. W. J. T. A. j. o. m. Brown, "Artificial intelligence in medical practice: the question to the answer?," vol. 131, no. 2, pp. 129-133, 2018.
- [12] K.-H. Yu, A. L. Beam, and I. S. J. N. b. e. Kohane, "Artificial intelligence in healthcare," vol. 2, no. 10, pp. 719-731, 2018.
- [13] A. Albahri and R. A. J. J. o. M. S. Hamid, "Role of biological Data Mining and Machine Learning Techniques in Detecting and

- Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," vol. 44, no. 7, 2020.
- [14] Z. Yang et al., "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," vol. 12, no. 3, p. 165, 2020.
- [15] Hameed Abdulkareem, K., Awad Mutlag, A., Musa Dinar, A., Frnda, J., Abed Mohammed, M., Hasan Zayr, F., Lakhan, A., Kadry, S., Ali Khattak, H. and Nedoma, J., 2022. Smart Healthcare System for Severity Prediction and Critical Tasks Management of COVID-19 Patients in IoT-Fog Computing Environments. *Computational Intelligence and Neuroscience*, 2022.
- [16] M. Alsalem et al., "Multiclass benchmarking framework for automated acute Leukaemia detection and classification based on BWM and group-VIKOR," *Journal of medical systems*, vol. 43, no. 7, p. 212, 2019.
- [17] O. Albahri et al., "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," 2020.
- [18] T. B. Alakus, I. J. C. Turkoglu, Solitons, and Fractals, "Comparison of deep learning approaches to predict COVID-19 infection," p. 110120, 2020.
- [19] M. A. Mohammed et al., "Benchmarking Methodology for Selection of Optimal COVID-19 Diagnostic Model Based on Entropy and TOPSIS Methods," in *IEEE Access*, vol. 8, pp. 99115-99131, 2020, doi: 10.1109/ACCESS.2020.2995597.
- [20] Abdulkareem, K.H., Al-Mhiqani, M.N., Dinar, A.M., Mohammed, M.A., Al-Imari, M.J., Al-Waisy, A.S., Alghawli, A.S. and Al-Qaness, M.A., 2022. MEF: Multidimensional Examination Framework for Prioritization of COVID-19 Severe Patients and Promote Precision Medicine Based on Hybrid Multi-Criteria Decision-Making Approaches. *Bioengineering*, 9(9), p.457.
- [21] Mohammed, M.A., Al-Khateeb, B., Yousif, M., Mostafa, S.A., Kadry, S., Abdulkareem, K.H. and Garcia-Zapirain, B., 2022. Novel crow swarm optimization algorithm and selection approach for optimal deep learning COVID-19 diagnostic model. *Computational intelligence and neuroscience*, 2022.
- [22] Saeed, M., Ahsan, M., Saeed, M.H., Rahman, A.U., Mehmood, A., Mohammed, M.A., Jaber, M.M. and Damaševičius, R., 2022. An optimized decision support model for COVID-19 diagnostics based on complex fuzzy hypersoft mapping. *Mathematics*, 10(14), p.2472.
- [23] Schwab, P., DuMont Schütte, A., Dietz, B. and Bauer, S., 2020. Clinical predictive models for COVID-19: systematic study. *Journal of medical Internet research*, 22(10), p.e21439.
- [24] Alyasseri, Z.A.A., Al-Betar, M.A., Doush, I.A., Awadallah, M.A., Abasi, A.K., Makhadmeh, S.N., Alomari, O.A., Abdulkareem, K.H., Adam, A., Damasevicius, R. and Mohammed, M.A., 2022. Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. *Expert systems*, 39(3), p.e12759.
- [25] Alballa, N. and Al-Turaiki, I., 2021. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in medicine unlocked*, 24, p.100564.