Detection and Classification of COVID-19 Using Machine Learning

Osama R. Shahin (poonkuntranphd2020@gmail.com)
Jouf University
Hamoud H. Alshammari
Jouf University
Ahmed I. Taloba
Jouf University
Rasha M. Abd El-Aziz
Jouf University

Research Article

Keywords: Detection and Classification, COVID-19, Machine Learning, virus, coronavirus, doctors

Posted Date: October 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-942284/v1

Detection and Classification of COVID-19 using Machine Learning

Osama R. Shahin [a], Hamoud H. Alshammari [b], Ahmed I. Taloba [a], Rasha M. Abd El-Aziz [a]

[a] Department of Computer Science, College of Science and Arts in Gurayat, Jouf University, Saudi Arabia

[b] Information Systems Department, College of Computer and information sciences, Sakaka, Jouf University, Saudi Arabia

poonkuntranphd2020@gmail.com

Abstract:

As people all over the world are vulnerable to be affected by the COVID-19 virus, the automatic detection of such a virus is an important concern. The paper aims to detect and classify coronavirus using machine learning. To spot and identify coronavirus in CT-Lung screening and Computer-Aided diagnosis (CAD) system is projected to distinguish and classifies the COVID-19. By utilizing the clinical specimens obtained from the corona-infected patients with the help of some machine learning techniques like Decision Tree, Support Vector Machine, K-means clustering, and Radial Basis Function. While some specialists believe that the RT-PCR test is the best option for diagnosing Covid-19 patients, others believe that CT scans of the lungs can be more accurate in diagnosing coronavirus infection, as well as being less expensive than the PCR test. The clinical specimens include serum specimens, respiratory secretions, and whole blood specimens. Overall, 15 factors are measured from these specimens as the result of the previous clinical examinations. The proposed CAD system consists of four phases starting with the CT lungs screening collection, followed by a pre-processing stage to enhance the appearance of the ground glass opacities (GGOs) nodules as they originally lock hazy with fainting contrast. A modified K-means algorithm will be used to detect and segment these regions. Finally, we will use the detected, infected areas that obtained in the detection phase with a scale of 50x50 and we will crop the solid false positives that seem to be GGOs as inputs and targets for the machine learning classifiers, here we will use a support vector machine (SVM) and Radial basis function (RBF). Moreover, a GUI application is developed which avoids the confusion of the doctors for getting the exact results by giving the 15 input factors obtained from the clinical specimens.

Introduction

The recent pneumonia pandemic, which started in early December 2019 around Wuhan Town in Hubei Province, China, is produced by a novel coronavirus (CoV) designated either by World Health Organization as "2019-nCoV" or "2019 novel coronavirus" or "COVID-19" (WHO). COVID-19 is a virus that can cause disease (Li et al. 2020). According to phylogenetic study using known whole genome sequences, bats are related with the COVID-19 viral reservoir, but the intermediate host(s) has yet to be discovered. It posed a global threat, similar to the Acute Respiratory Distress Syndrome (ARDS) and Severe Acute Respiratory Syndrome (SARS), both

of which are Coronavirus-related diseases (Wang et al. 2020). The WHO acknowledged the COVID-19 epidemic a "public health emergency," stating that the virus spreads to healthy people who come into contact with virus-infected people through the respiratory tract. It can also be transferred in a variety of different ways that experts are still trying to figure out. Infected people's symptoms, such as fever, cough, and pneumonia, will appear in 2 to 14 days.

COVID-19, a revolutionary coronavirus, revolutionized the world's healthcare system. In addition to healthcare, global economics, education, and transportation have all been altered (Fong, Dey, and Chaki 2021). This virus disease can cause serious respiratory sickness, but with correct treatment, it can be healed. However, the virus's most hazardous side effects include human-to-human transmission and community proliferation. In cluster cases, a prediction based on the artificial intelligence (AI) is possible to detect them using this method.. Additionally, previous clinical data can be used to make this prediction. AI can work in a fashion that is similar to the human brain. Furthermore, AI can comprehend and portray the progress of the COVID-19 vaccine development (Kondziolka, Couldwell, and Rutka 2020). The present patient tracking, screening, analyzing, and predicting should be done for an accurate prediction of COVID-19 instances, which can help in the future prediction of infected patients. AI is now routinely utilized to find new compounds for the development of a COVID-19 assistance. Many studies are being conducted to find new treatments for curing the disease, as well as computing to detect disease-affected persons using medical image processing of CT scans and X-ray pictures.

The sickness is confirmed using the reverse-transcription polymerase chain reaction (RT-PCR). The RT-PCR sensitivity is insufficient to detect disease because it lacks the sensitivity required to proceed with the treatment of suspected patients (Sun et al. 2020). Computed tomography is utilized to discover a certain distinctive manifestation in the lungs. This CT can be utilized to perceive the case since COVID-19 affects the lungs. CT-image screening can be utilized to identify the disease or COVID-19 virus at an early stage. COVID-19's CT picture demonstrates the pneumonia disease's resemblance (Zebin and Rezvy 2021). Once one person gets into close touch with someone who is afflicted, the virus transmits to the other. It can spread through when an infected individual breathes, coughs, or sneezes, their nose and mouth is infected. The virus quickly spread throughout the infected surface. The conventional diagnostic procedure was employed to detect the virus. The nucleic acid is amplified from a nasopharyngeal swab using RT-PCR, transcription-mediated amplification (TMA), or loo-mediated isothermal amplification (Khalifa et al. 2020). To stop the corona virus spreading, many precautions were taken.

The paper goals to establish effectiveness of big data and AI to use these technologies to combat laboratory results derived from clinical specimens of coronavirus suspects, as well as to analyze recent solutions (Abbas, Abdelsamea, and Gaber 2021). Computer-aided diagnosis (CAD) is a medical procedure that assists clinicians in explaining the clinical specimens of coronavirus patients (Yuan et al. 2020). The abdominal problem is discovered while looking at the patient's CT chest imaging. When the condition becomes more advanced, the patient may have breathing

issues, heart damage, and other infections. This can result in death, so it's critical to catch the corona early on.

The contribution of the proposed method:

- The proposed method is held in detecting and classifying the COVID-19 virus. The basic method used in this system is a machine learning method.
- The CT-lung screening approach is used to do this. The respiratory system is the first organ to be impacted by the covid-19 virus. As a result, a scan and diagnosis of the lungs or respiratory system are required. Also, using clinical samples to develop a machine-learning model for detecting the Coronavirus in the patients.
- Building a CAD system that collects data from COVID-19 patients or suspects and determines if the patient is infected or not. To enhance accuracy and access to data in less time, modern machine learning methods are applied.

Related Work

(Al-antari et al. 2021) proposed CAD system was assessed using five-fold tests for the multi-class prediction issue utilizing two independent databases of chest X-ray images, COVID-19 and ChestX-ray8. The suggested CAD system was trained using an identified training set of chest X-ray images. The proposed CAD predictor was employed to determine and categorize regions on entire X-ray images with COVID-19-related lesions, with overall detection and diagnosis accuracy of 96.31 percent and 97.40 percent, accordingly. With such a mean intersection over union (IOU) of better than 90%, the most of test photographs from COVID-19 and other respiratory ailment patients were accurately predicted. Deep learning regularizes data balancing and enhancement improved COVID-19 diagnosis efficiency by 6.64 percent and 12.17 percent, respectively, in terms of overall accuracy and F1-score. A diagnosis based on individual chest X-ray photographs can be made in 0.0093 seconds using the specified CAD method. As a result, the Design used in this study can forecast at 108 frames a second (FPS), which is close to real-time. Using proposed deep learning CAD system, COVID-19 may be reliably identified from other respiratory illnesses. In the real world, the suggested learning algorithm looks to be a reliable tool for supporting health care systems, consumers, and physicians.

(McKee et al. 2015) The purpose of this research was to determine whether ACR Lung-RADS affects the rate of false-negative and positive findings in a clinical CT lung screening test. The goal of this review was to determine how ACR Lung-RADS affects the rate of false-negative and positive findings in a clinical CT lung screening. A number of 2,180 high-risk individuals had a baseline CT lung examination during the research period, with 577 patients receiving no clinical follow-up. The overall positive rate for ACR Lung-RADS was reduced from 27.6% to 10.6%. There are still no false negatives among the 152 patients classified as benign after a 12-month follow-up. Using ACR Lung-RADS increased the predictive value for detecting cancer from 6.9% to 17.3 percent in 1,603 patients with follow-up. The introduction of ACR Lung-RADS

increased our CT lung screening group's positive predictive value by a factor of 2.5, to 17.3 percent, while lowering the number of false-negative tests.

(Zou et al. 2021) The purpose of this study was to determine what factors impact the frequency of positive RT-PCR results. Using a retrospective analysis, we looked at the clinical information of recurring positive coronavirus disease 2019 (COVID-19) patients in multiple medical facilities in Wuhan. Based on their RT-PCR results, patients are separated into two groups: recurrent positives and nonrecurrent positives (non-RPos group). Clinical characteristics, updated content, and antibody titers was placed in two groups. They used AI-assisted chest increased computed tomography (HRCT) equipment to examine pulmonary inflammatory exudation and assess the size of lung sections with varied densities. This study included 122 COVID-19 participants. In terms of age, gender, past diseases, clinical manifestations, clinical classification, clinical history, medication regimens, or serum-specific antibodies, there are still no significant differences between the two groups. COVID-19 recurrence is associated with subpleural exudation towards the lung periphery and severe respiratory failure at discharge.

(Irmak 2021) provides a unique COVID-19 illness severity classification approach based on a convolutional neural network (CNN). Using chest X-ray images as input, an automated CNN model is constructed and proposed to split COVID-19 patients into four severity classes: mild, moderate, severe, and critical with an average accuracy of 95.52 percent. Experiments on a sufficiently large number of chest X-ray images illustrate the efficacy of the CNN model constructed with the suggested framework. It's the first COVID-19 injury severity evaluation study comprising 4 stages, using a sufficient huge number on X-ray data sets and a CNN with virtually all hyper-parameters dynamically adjusted mostly by variable selection optimization, as far as the researcher is aware.

(Mazzilli et al. 2021) Millions of individuals around the world are affected by the ongoing COVID-19 pandemic. Chest computed tomography (CT) is the most widely used imaging modality, and it is critical for patient diagnosis and treatment. The lungs of COVID-19 patients were described using an automated methodology based on individual adapted Hounsfield unit (HU) thresholds. The HU-density calibration curve's impact on inter-scanner variability was explored. Inter-scanner variability was found to be insignificant. Individual thresholds th1 had median values of 768, 780, and 798 HU for the three techniques, respectively. In comparison to the other two methods, the maximum gradient of the data had a substantially lower median value. Three factors of our electorate were quantified using a millimeter gradient on the data method; aerated, intermediate and consolidation components had median values of 793 499 cm3, 914 291 cm3, and 126 111 cm3, respectively; the first peak had an average value of 853 56 HU, and the second peak had an average value of 854 56 HU.

Proposed system

This system proposed detection and classification of CT-lungs screening and clinical specimens of COVID-19 using machine learning methods. The clinical procedure for infection diagnosis in

the COVID-19 clinical specimen samples takes more time and effort by the doctors and also for the patients. This method is used for early infectious disease detection. Then the process of image collection, starting with a pre-processing stage to improve the appearance of the ground glass opacities (GGOs) nodules, which were previously fuzzy with fading contrast, the process of gathering CT images and creating a classifier model will take place (Aminisefat and Saravani 2020). To detect and classify the CT- Lungs screening has been proposed in two-phase the one is the building of classifier model and the other is testing a new CT image. Then detect the COVID-19 in clinical specimens is supervised by machine learning techniques such as Naive Bayes Classification, Decision Trees, Support vector machines, Radial basis function, and K-means clustering are used and trained on the dataset to effectively distinguish between infected and non-infected COVID-19 cases in clinical test samples. The following sections explain the detection and classification of COVID-19 are shown in figure 1.

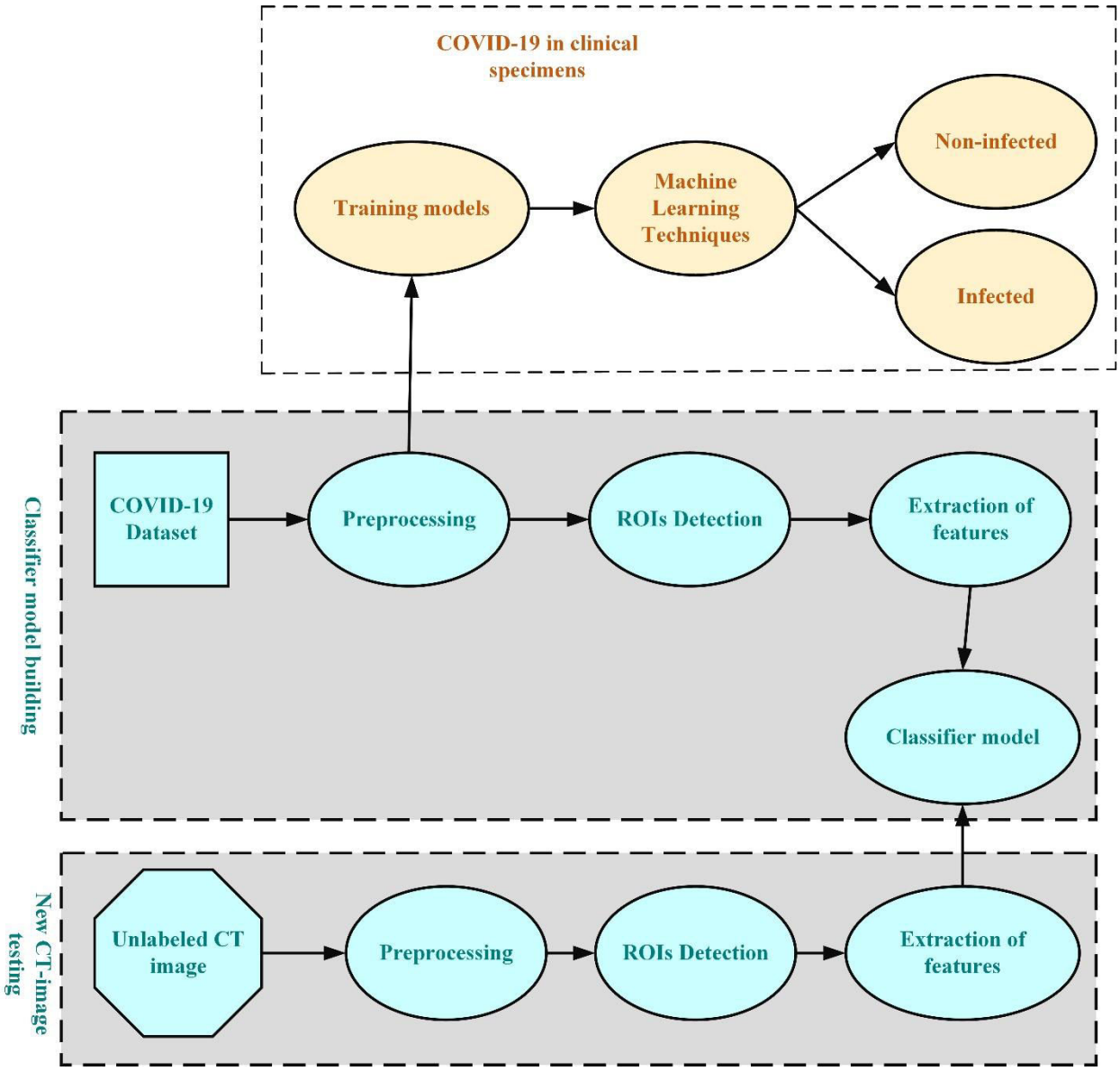

Figure 1 proposed model of Clinical specimens and CT-lungs screening

Several machine-learning models based on several supervised learning approaches such as Naive Bayes Classification, Decision Trees, Support vector machines, Radial basis function, and K-means clustering are studied with the collected dataset in the third phase. Figure 1 also depicts the proposed CAD design.

COVID-19 dataset:

As part of CT-lung screening, the data set is used to detect affected and normal COVID-19 virus data. This information provides a clear automatic segmentation scheme and quantifies anomalous CT models. The collection of the dataset is considered a difficult task because they require a large number of ethical and privacy of the hospital system (Zhao et al. 2020). Based on ethics and law, the appropriate commissions accepted this dataset. Different forms of the disease and scanning methods are included in the collection. The patient will be categorized as corona or not corona based on the given dataset. If a dataset has 100 samples, the corona disease will infect at least half of them. The entire data set is gathered, categorized according to the ailment, and then processed to the next phase. They preserved and secured the data for subsequent use. The clinical specimens for SARS-CoV-2 testing are given as follows:

- Whole blood specimens.
- Respiratory secretion.
- Serum Specimens.

Preprocessing

This step involved classifier model building. The sample is separated from the dataset is further processed to the preprocessing section. This section contains the CT scan slices from which the data is gathered. 2D and 3D sample slices are included in the slides. The facility and medical images are dimensioned and normalized as part of the preprocessing step. In this step, the medical image is diagnosed and presented. This leads to the following stage, which is the identification of ROIs. On clinical specimen preprocessing, the unstructured data must be refined before proceeding to the next detection stage. Several elements must be taken into account during the preprocessing stage. Samples with insufficient or ambiguous data should be discarded. To improve accuracy, the unstructured texts in the dataset will be preprocessed for punctuation, lemmatizations, symbols, stop words, and URLs.

ROIs detection

Following the preprocessing test, the Regions of Interest (ROIs) are processed. The RGB color is employed in this technique to accomplish the moving window to the segment region of interest (ROIs) (Jiang, Wang, and Liu 2015). This focuses primarily on the location of the moving thing, with the consequences of the moving object being rejected. In this detection, the color

concentration is more crucial. The colors red, blue, and green are utilized to indicate the system's detection. These colors could be used to distinguish the areas that are influenced by the corona.

Feature extraction

Each layer in the x-ray is extracted and then a systematic result is created as part of the feature-based part of the capsules structure. It is a procedure in which one image is categorized and compared to other images to obtain a result.

Classification and Detection

In CT imaging of a corona patient's lungs, infected areas appear as Solid nodules and Ground Glass Opacity (GGO) (Kim et al. 2009). The enhanced attenuation without obscuration of the underlying arteries and bronchi in Ground Glass Opacity (GGO) nodules obscures the whole lung parenchyma inside it. GGO nodules are categorized as follows:

- A GGO-containing nodule (part-solid nodule)
- A GGO-containing nodule with pure localized GGO (non-solid nodule)
- The nodular ground-glass Opacity, also known as the subtle nodule, is a technique that is utilized in clinical practice, computed tomography (CT) scanning and image screening were widely employed. The lungs scanning system employs this technology.

In medicine, computer-aided detection (CAD) or computer-aided diagnostics (CAD) are tools that aid clinicians in deciphering medical images. The radiologist must analyze and estimate a huge amount of data in minimum amount of period using the imaging modalities including X-ray, MRI, Ultrasound, and CT diagnoses. CAD systems aid in the scanning of digital images to detect various lung locations, such as probable infections. Computer-aided diagnosis is described as the process of a person making a decision about his or her medical health and diagnosing that process (CAD). This computer output is used to process a version of an original image.

The suggested system's major phase is the categorization model. This system must be able to distinguish between positive and negative images of Corona virus-infected lungs and other normal images. Numerous machine-learning algorithms may be used to create such a model. We'll use two classifiers in this case. Radial basis function (RBF) networks, for example, have 3 layers: an indoor layer, hidden layer, and outside layer. It is critical to set suitable initial states for RBF networks since it is becoming increasingly popular in neural networks with various applications. It is probably the most competent to the multi-layered perceptron. The vector supporting machine will be utilized as the second classifier (SVM). SVM is also a good classifier for digital picture classification, especially when the categorization is based on colors or specific features.

Classifier Model

The classifier model is the last and most important step in the classification process. The image extracted through feature extraction is categorized in this method based on the disease that has

been diagnosed. The same process is repeated when a new CT picture is tested, eventually leading to an unlabeled CT image. The person who has been afflicted by the covid-19 has been identified as a result.

Flowchart for COVID-19 detection

The workflow for detection of COVID-19 is shown in Figure 2. The first step is to perform the Covid-19 test, which is the first stage of the test. Continuing, a CT-lung screening is performed (Zhang, Chu, and Zhao 2020). In CT-lung screening, the person undergoes a lung test, which is then processed by preprocessing and feature extraction. The CT image is retrieved and then diagnosed through a series of techniques known as feature extraction. The patients are then categorized according to their ailment and tested again. The COVID-19 test is then performed based on this information, and the results are analyzed. There are two types of results: normal and covid positive.

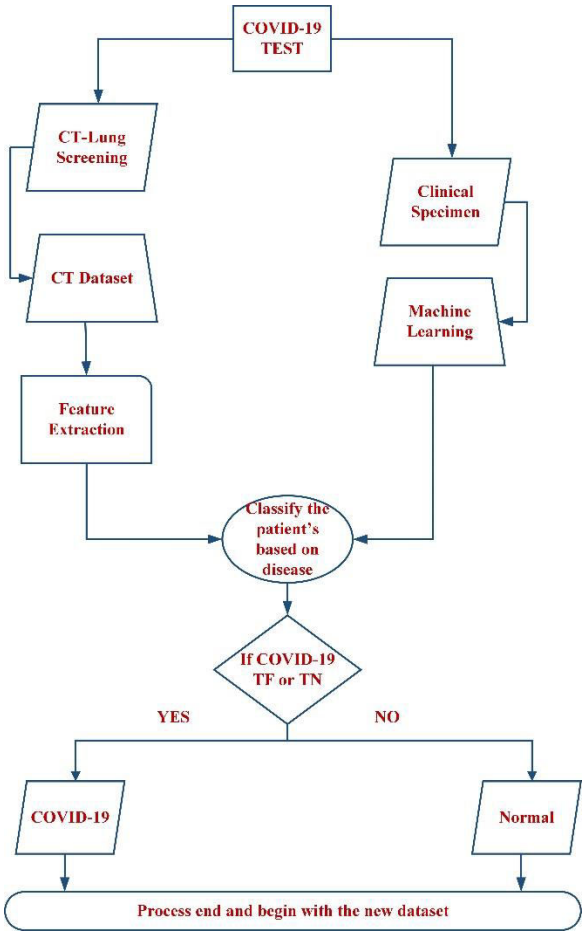

Figure 2 Flowchart for the COVID-19 detection

Training Models

The generated dataset is analyzed using different machine learning models in this step to check its accuracy and compare it to the accuracies of each model to determine the optimal model for real-time COVID-19 detection and classification.

Decision Tree

A Decision Tree (DT) is a tree-like structure used to model decisions and their outcomes. The internal nodes in a DT denote the attribute test, whereas the branch denotes the outcome. The leaf nodes represent the class labels (Kumar Dubey et al. 2021). Because it does not require domain-specific expertise to develop, DT could be beneficial in a variety of contexts. Classification and Regression Trees are other names for DTs (CART). The Decision Tree (DT) is a classification and prediction machine learning algorithm. The decision tree algorithm is straightforward to create and modify. In terms of efficiency and processing speed, the decision tree algorithm performs admirably. As a result, as compared to the principle of unit classification, the accuracy of the decision tree machine learning algorithm is higher.

DT and LR are two common classification methods. Algorithm 3 depicts the key steps taken by the DT and LR classifiers for predicting the COVID-19 artificial individuals.

Algorithm 1: DT and LR for detection of positive COVID-19 patients

Input: The following is a feature ranking.

Output: confusion matrix, classification report, accuracy

Process:

i. Using the StandardScaler ( ) function, standardize the selected characteristics.
ii. Applying DT to the specified features using the DT Classifier (criterion='entropy', max_depth=5, random_state=0) function with some parameters.
iii. Select features are used to train the model.
iv. K-fold cv specifications: thresh=0.5, k_fold_seed=13, n_folds=10.
v. Using the test dataset, forecast the outcome.
vi. Predict the outcome using the test dataset.
vii. To evaluate FN, FP, TN, and TP by using confusion_matrix()
viii. Calculate recall, precision, and F1 score with classification report () function.

The information such as domain names is used in the decision tree algorithm in figure 1. Furthermore, this algorithm is capable of selecting the most unjust text data from a large number of raw texts, as well as managing noisy data. As a result, it is thought that using a machine learning technique like the decision tree is a superior but reasonable choice for achieving the goal of data being affected by multiple features via an information grouping mechanism. The decision tree algorithm includes a tree-like flow structure in figure 3, with hubs inside that, represent a property test, branches that represent the test result, and stems that classify the forecast.

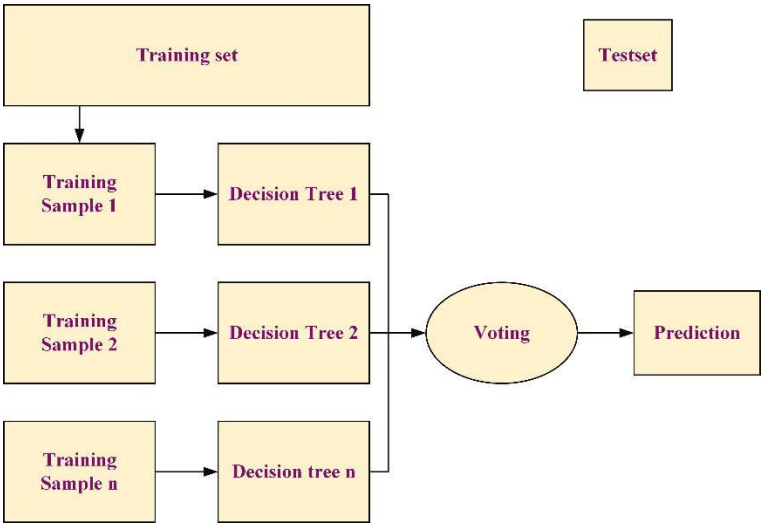

Figure 3 Decision Tree Algorithm

The most essential concern in using the decision tree method is determining the optimal size of the last tree. The over-fitting problem that emerges owing to the larger tree and the under-fitting problem that arises due to the smaller tree are the two key issues that arise when using a decision tree in a machine learning technique. Pruning a decision tree is the process of deleting data that is less important than the test data presented. The removal of this data during processing reduces the size of the data and improves the accuracy of the outcome. As a result, the pruning concept eliminates the categorization approach's complexity.

The decision tree is a predictive analysis approach that is supposed to work by creating a category based on the dataset it was fed with, which then creates its sub-categories, which again creates its sub-categories, and so on until the last node is created, which gives us the desired result or the programmed terminates. The model's training dataset will be used to generate the prediction result. With the increasing length and depth of the decision tree, the complexity of its execution rises. The decision tree approach algorithm in algorithm 2.

Algorithm 2: Decision Tree

Assume No. of Samples = S;

Data Points = p;

Target Inputs = q;

No. of Leaves = Y_s;

Tree_Depth = T;

Criterion = R;

for y in test size;

```
do
    for a in R do
        test_size = p_test and q_test;
        train_size = p_train and q_train;
        for e < p do
            Call function DT;
             For f < T do
                Evaluate the best_spilt;
                Evaluate class (r);
                Ys++;
                Node, (r,R);
                Return
                    Predicted_class (r);
            Calculate the Accuracy;
```

Naïve Bayes

Random forest is a classification and regression analysis algorithm that uses supervised learning. Tin Kam Ho used the random subspace method to build the algorithm. An ensemble of multiple decision trees, the random forest is composed. Based on the taking of majority votes in the tree class, each tree spits out a class prediction (Abdulkareem et al. 2020). The random forest's basic steps are to take a random sample from dataset and build a decision tree from each tree to produce an estimate. Vote on the prediction tree's final prediction and choose the one with the most votes.

The Bayes theorem is used to classify data, and the Naïve Bayes algorithm is based on it. This methodology was first utilized for text categorization in the 1960s under the way of text retrieval problems (Imad et al. 2020). The Bayes theorem allows you to compute the posterior probability of $P\left(\frac{x}{y}\right)$ by using p(x), p(y), and $P\left(\frac{y}{x}\right)$, as shown in equation 1.

$$P\left(\frac{x}{y}\right) = \frac{P\left(\frac{y}{x}\right)p(x)}{p(y)} \qquad (1)$$

In equation (1) presents a posterior probability of p(z/a) in the case of a specific predictor (z is a target and a is the attribute). The prior probability of the class and predictor, respectively, is p(z)

and p(a). The probability of a particular class is given by p(a/z). figure 4 shows the working of the Naïve Bayes Algorithm.

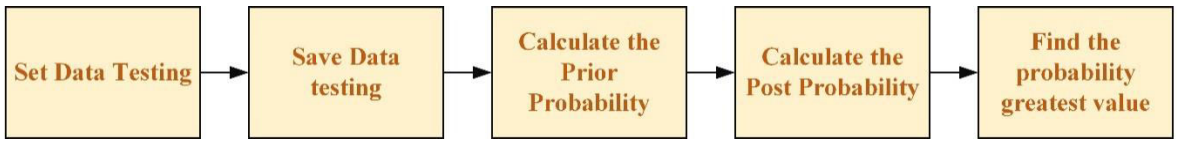

Figure 4 Naïve Bayes Algorithm working.

To determine the probability of a positive result in sample A with tokens {O_1,..., O_n, it is necessary to integrate tokens with varied positive rates to compute the overall rate of COVID-19 positive samples. To accomplish the categorization, we calculated the product of the infected rate for a single token and compared it to the product of the non-infected rate for that token directly.

If the total positive rate of the product, P [D], is greater than the negative rate of the product, N [D], the test sample is categorized as positive. As shown in Algorithm 3, the above equation is used in the classification of Naïve Bayes strategy for the COVID-19 organization.

Algorithm 3: Naïve Bayes

Input: COVID-19 dataset

Calculate the likelihood of each element by breaking it down into its constituent tokens.

$$D[O] = \frac{R_{infected}(O)}{R_{infected}(O) + R_{non-infected}(O)}$$

Store the values of infected in the database;

For each patient data D do

While (D not end) do

Image sample for following token, O_x;

Query folder for the infectious samples D(O_x);

Calculate the collected samples probability, P [D] and N[D];

Calculate the number of samples by: V[D]=h(D[D], J[D]);

$$V[D] = \frac{V+P[D]-N[D]}{2}$$

If V[D] > threshold:

Infected;

else

Non-infected;

End if

End while

End for

Return

Last classification (Infected/Non-infected);

end

Support Vector Machine

Detecting Unpredictable data makes determining COVID-19 from symptoms difficult. No suitable data set may therefore be used as a standard. According to the findings, the majority of COVID-19-infected persons were hospitalized with viral fever, respiratory infection, and trouble breathing (Guhathakurata et al. 2021). Once infected with COVID-19, Those with high blood pressure, heart problems, and a rapid pulse rate swiftly proceed to the next level. If the virus progresses to acute respiratory illness syndrome, it can cause respiratory failure, septic shock, and acute respiratory disease syndrome (ARDS). Using our proposed method, we can determine whether or not someone is infected based on their symptoms. Mildly infected, severely infected, and not infected are three classifications for the outcome, i.e., infected status. The numerical values have been mapped to the classes as follows in figure 5:

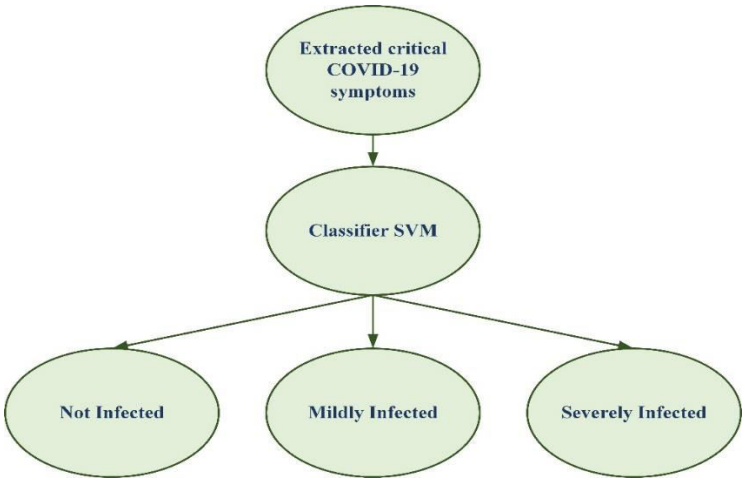

Figure 5 COVID-19 Detection Methodology

Not Infected:

Cases labeled as not infected only show signs of a single sickness, which is completely normal for any human being. People can get a moderate fever and a dry cough from certain illnesses,

such as the common cold, but such symptoms alone aren't enough to rule out COVID-19 infection.

Mildly Infected:

This classification denotes the symptoms do not conclusively indicate COVID-19, then they lead to the serious, if need the precautions are didn’t taken. In circumstances when the patient has a mild fever and a mild respiratory difficulty, this can speculate that patient is diseased with the COVID-19.

Severely Infected:

COVID-19 has shown favorable outcomes in the majority of patients with more than two or three symptoms, each of which is over the normal range. A high fever, rapid breathing, and acute respiratory syndrome are all signs of a serious illness for the patient. Next, the dataset is given to the SVM Classification model.

SVM was chosen for this problem because it employs the kernel method to transfer low-dimensional input space to high-dimensional space, effectively translating a non-separable issue into the separable issues. A train set has been created from the dataset. The SVM classifier uses a hyper-plane to linearly separate the data using the linear kernel. Parallel hyper-planes separate each data class, ensuring that the distance between them is as vast as possible. Detecting COVID-19 is a high-priority situation, hence we are looking for a hyper - plane with a narrower margin of error to better reliably classify the infected classes. The SVM working principle is in figure 6.

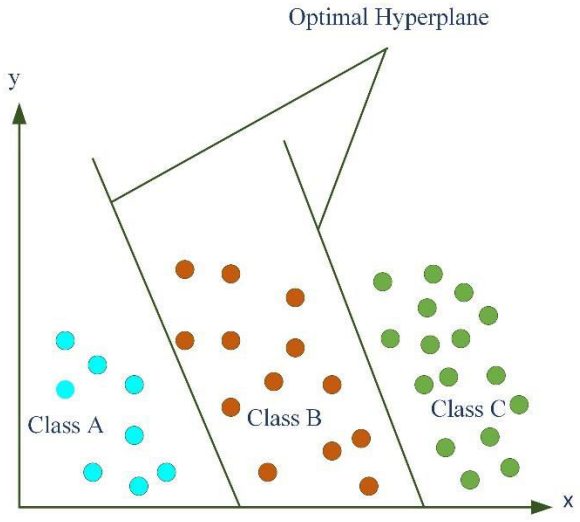

Figure 6 SVM Working Principle

For the categorization of clinical samples, the C-SVM classification algorithm is used in this research. The C in C-SVM stands for the cost parameter, which controls the modeling error that happens when the function is an overfitted to sets of data points, resulting in a mistake. To begin

the process of training, consider that a collection of data is being trained, with (C, γ) as a hypothetical merger parameter capable of developing a SVM superior classifier. The Grid exploration in γ and C parameters is a practical strategy that is commonly used in the SVM classifier to find this fusion limitation. In the grid-search, the k-fold rotation estimation is used to find the SVM classifier with best rotation approximation accuracy prediction. Algorithm 3 is the SVM Classifier's algorithm.

Algorithm 4: Support Vector Machine (SVM)

Input Clinical Sample data x for classification:

Training set = T;

Kernel function = $\{C_1, C_2, \ldots C_{num}\}$ and $\{\gamma_1, \gamma_2, \ldots \gamma_{num}\}$;

Number of nearest neighbors = k;

for $x = 1$ to number:

set $C = C_i$;

for $y = 1$ to q:

set $\gamma = \gamma$;

Create a trained SVM classifier $f(a)$ via the fusion parameter (C, γ);

if ($f(a)$ is the discriminant function formed first):

$f(a)$ is the most perfect SVM classifier $f*(a)$;

else

Compare the classifier $f(a)$ and present greatest SVM classifier $f*(a)$ by k-fold cross-validation;

Maintain the classifier with the better accuracy;

end if

end for

end for

return

Classification of Result finally (Infected/Non-infected);

end

Radial Basis Function

It's an ANN that utilizes activation functions of radial basis, as illustrated in the Equation. The RBFNN is a feed-forward neural network having a 3 levels (Dhamodharavadhani, Rathipriya, and Chatterjee 2020). While the first layer transmits only the input signal, while the second level uses non-linear Gaussian functions. Lastly, the third layer includes the Gaussian linear outputs. During training, only its tap weights between the hidden layer and the output layer were changed.

$$f(a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2)$$

The Gaussian function f(a) is used to approximate a function in equation 2. The real values are represented by a. The input a is used to determine the function's dimensional parameters. In generalization issues, the radial basis function (RBF) network is an extensively used as Artificial Neural Network (ANN). The RBF network differs from other neural networks in that it learns more quickly and has universal approximate. It is made up of 3 layers: the input, the output, and the hidden, of all which are associated by a feed-forward network. Each layer has a distinct purpose.

When the error reaches the target value of 0.01 or the total iteration of the training reaches 500 times, the RBF model training can be ended. The RBF is chosen in such a way that the hidden layer of the RBF must have a total of 10 nodes. The Gaussian function is employed as the transfer function in the computational units. The RBF network is quick and efficient, taking less time to complete the training. Figure 7 shows the working of the Radial Basis Function.

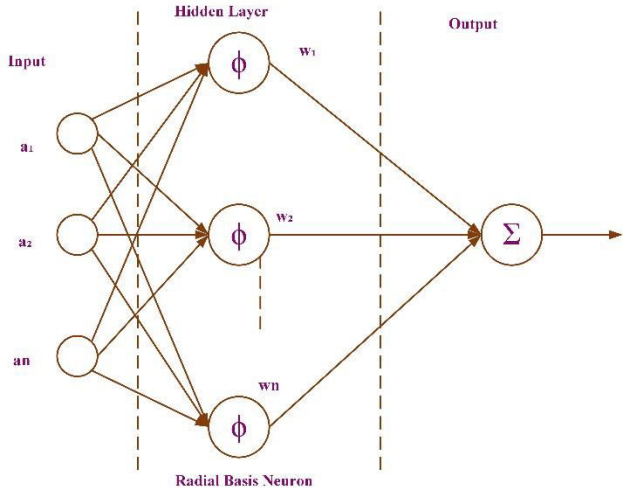

Figure 7 Radial Basis Function Working

The RBF has two inputs, which are the coordinates of a point, and a single output in the approximation situation of a 2-variable function. The phrase is used to calculate the RBF output with the input A= $[a_1, a_2]^T$ in equation (3).

$$u(a) = \sum_{t=1}^{n_{RBF}} w_t \varphi_t (a) \qquad (3)$$

where, n_{BRF} = Number of RBFs,

w_t = Weight of the t^{th} neuron,

A = Input vector,

$\varphi_t(X)$ = Value of t^{th} RBF at this point.

A Gaussian function is used as an RBF, which, in its two-dimensional case, is written as:

The 2-D expressional for the Gaussian function is expressed in (4).

$$\varphi(A) = exp\left(-\frac{\|A - C\|^2}{2x^2}\right) \qquad (4)$$

were, $C = [c_1, c_2]^T$ → vector of coordinates of the RBF center,

x = width,

The Euclidean norm is given as (5)

$$\|A - C\| = \sqrt{(a_1 - c_1)^2 - (a_2 - c_2)^2} \qquad (5)$$

Errors can be minimized by using equation (6).

$$H = \frac{1}{2}\sum_{v=1}^{n} e_v^2 = \frac{1}{2}\sum_{v=1}^{n}\left(u(P_v - T_v)\right)^2 \qquad (6)$$

where n = number of test points,

e_v = solution error at j^{th} test point,

P_v = coordinates of j^{th} test point, in case of approximation of a 2-variable function $P_v = \left[p_{v_1}, p_{v_2}\right]^T$,

$u(P_v) = u(a)$ at the j^{th} test point,

T_v = target value at j^{th} test point,

multiplier $\frac{1}{2}$ is used for simplification.

K-means Clustering

The k-means technique is used to classify or cluster the data collected from various countries. The dataset was clustered using a change of methods, it including a Density-based Spatial Clustering of Application with the Noise (DBSCAN), k-means, and the Hierarchical Clustering Algorithm (Nour, Cömert, and Polat 2020). The k means the technique is widely utilized because of its distance-based, quick processing, and linear calculation. The k-mean can be determined using the steps below.

There are two stages to the K-means clustering technique. In the first phase, an average linkage technique is used to perform a preliminary categorization. The clustering centres obtained in the first phase as the initial clustering centre are employed in the typical K-means approach in subsequent classifications in the second phase. The expected number of clusters is given as K and is determined by the dataset's characteristics.

Algorithm 5: K-means Algorithm

Step1: Choose the number of ‘k' required to identify the object's spatial representation.

Step2: In the primary group of centroids, this must be represented.

Step3: The data point is calculated between the points of respective centroids.

Step4: The point that is closest to the center is labeled.

Step5: The centroid group is recalculated based on the classified point.

Step6: Steps 2 and 3 are performed until the centroids no longer change.

To detect and classify the covid-19 data set, the k-means are calculated. When compared to the other algorithms, this is the most effective strategy.

First, initialize the clustering centers from the first phase in the second phase of the K-means clustering approach. The samples are then classified into clusters based on how similar their initial centers. Then, for successive iteration, compute the average value of the clusters as centers. These steps must be repeated until the clusters have been altered. The working of K-Means clustering is demonstrated in figure 8.

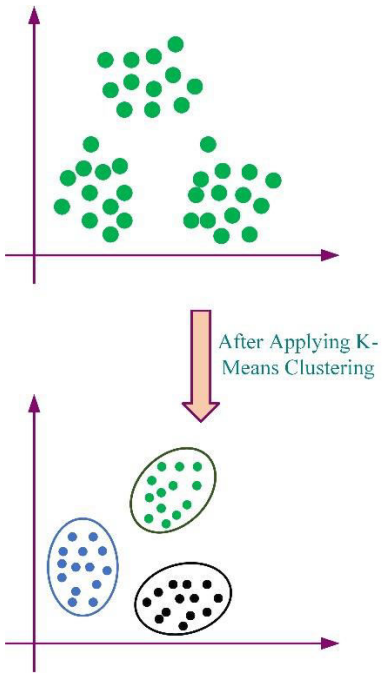

Figure 8 K-Means clustering Working

"The k-NN" algorithm is one of the most extensively used and basic classification algorithms. The kNN algorithm is slow, nonparametric method. Let's try to comprehend the idea of passive learning, we will find that it does not have a training period, unlike eager learning. When wish to make an estimate, it searches the entire dataset for the closest neighbors. A k value is determined during the algorithm study. The number of elements to be examined is represented by the k value. As soon as a value is obtained, the range between the k closest element is calculated. It is usual to utilize the Euclidean function in distance calculations. The Euclidean function can be replaced by the City Block, Minkowski, and Chebyshev functions. This is followed by sorting and assigning the incoming value to the relevant class.

Result and Discussion

The classifier decision was used to determine the performance measure. The classifier determines whether or not the person is affected by the disease. This classification can be separated into four categories, the first two of which are infected and normal, and the other two of which are due to a testing error. The first section is called Test Positive and Negative, while the second section is called False Positive and False Negative. The test positive and negative results indicate whether or not a person is impacted by covid. The false-positive indicates that the person is normal, but the test indicates that the person is not, while the false negative indicates that the person is impacted by the covid-19, but the test indicates that the person is not. The sensitivity measurement can be seen in the false-positive and false-negative results.

Confusion Matrix

Metrics	Test Result (Positive or Negative)	
Actual biopsy (True)	True Positive (TP)	False Positive (FN)
Actual biopsy (False)	False Negative (FP)	True Negative (TN)

Table 1 Confusion Matrix

The performance of the proposed COVID-19 detection and classification algorithms can be assessed using a variety of performance indicators. To control the presentation of the suggested structure, a Confusion Matrix is produced. Table 1 shows the Confusion Matrix. Four metrics can be used to specify the performance metrics that will be used to evaluate the classification model: accuracy, correctness, recall, and F1 score.

Accuracy

The accuracy of a proposed model can be assessed by expressing the average of the True values in the obtained results from a given dataset of clinical specimens as the proportion of the sum of True Positives (TP) and Negatives (TN) to the total of True Positives (TP) and Negatives (TN) and False Positives (FP) and Negatives (FN), as shown in (7)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \qquad (7)$$

Precision

The ratio of True Positives (TP) to the sum of False Positives (FP) and True Positives acquired from a particular dataset of clinical specimens is used to evaluate the performance metric precision is expressed in (8)

$$Precision = \frac{TP}{TP + FP} \qquad (8)$$

Recall

The metric for measuring performance the ratio of True Positives (TP) to the sum of False Negatives (FN) and True Positives derived from a dataset of clinical specimens is used to recall determine in (9).

$$Recall = \frac{TP}{TP + FN} \qquad (9)$$

F1-Score

Precision and Recall metrics are used to calculate the F-measure, F_. The measure F1-score is determined by the value of F_. The Precision and Recall on Harmonic Mean as indicated in the F1-score is expressed in (11).

$$F_\beta = \frac{(1+\beta^2)(Precision * Recall)}{\left(\beta^2 * (Precision + Recall)\right)} \quad (10)$$

Substitute $\beta = 1$, after simplification,

$$F_\beta = \frac{2 * (Precision * Recall)}{1 * Precision + Recall} \quad (11)$$

Conclusion:

The goal of the research is to develop the technique for detecting and classifying the covid-19 virus. The machine learning method is used to recognize and classify objects in this manner. The COVID-19's effects astounded the globe because there were no specific treatments or vaccines available at the time of the virus's emergence. Several studies are being conducted to overcome this lethal disease caused by Coronaviruses. These clinical specimens were subjected to genetic, serological, and biochemical testing. For further processing, several characteristics are collected from clinical samples, and also the CT-lung screening speeds up the process and aids in identifying people who have been infected with the coronavirus. Based on this, the covid test is performed, the genuine positive, negative, and false positive, negative are separated, and the accuracy, precision, recall, and F1-factor are calculated. Because the prediction rate is so high, it's much easier to locate those who are afflicted. To lower the death rate. As a result, utilizing machine learning to detect and categorize COVID-19 in CT lungs screening is deemed the best way for detecting and classifying the virus. The efficiency and accuracy of the suggested model can be enhanced by increasing the number of samples used. For a better outcome, more feature engineering is required, and deep learning can be applied in the future.

Funding

There is no funding.

Conflict of Interest

This paper has not communicated anywhere till this moment, now only it is communicated to your esteemed journal for the publication with the knowledge of all co-authors.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent

All authors have seen the manuscript and approved to submit it to the journal.

Author contributions

All authors have seen the manuscript and approved to submit it to the journal.

References

1. Abbas, Asmaa, Mohammed M. Abdelsamea, and Mohamed Medhat Gaber. 2021. "Classification of COVID-19 in Chest X-Ray Images Using DeTraC Deep Convolutional Neural Network." *Applied Intelligence* 51 (2): 854–64. https://doi.org/10.1007/s10489-020-01829-7.
2. Abdulkareem, Nasiba M, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, and Dathar A Hasan. 2020. "COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms," 6.
3. Al-antari, Mugahed A., Cam-Hao Hua, Jaehun Bang, and Sungyoung Lee. 2021. "'Fast Deep Learning Computer-Aided Diagnosis of COVID-19 Based on Digital Chest x-Ray Images.'" *Applied Intelligence* 51 (5): 2890–2907. https://doi.org/10.1007/s10489-020-02076-6.
4. Aminisefat, Alireza, and Khadijeh Saravani. 2020. "A Case Study of Hypertension and COVID-19." *Gene, Cell and Tissue* 7 (3).
5. Dhamodharavadhani, S, R Rathipriya, and Jyotir Moy Chatterjee. 2020. "COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models." *Frontiers in Public Health* 8 (August): 441. https://doi.org/10.3389/fpubh.2020.00441.
6. Fong, Simon James, Nilanjan Dey, and Jyotismita Chaki. 2021. "An Introduction to COVID-19." In *Artificial Intelligence for Coronavirus Outbreak*, by Simon James Fong, Nilanjan Dey, and Jyotismita Chaki, 1–22. SpringerBriefs in Applied Sciences and Technology. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-5936-5_1.
7. Guhathakurata, Soham, Souvik Kundu, Arpita Chakraborty, and Jyoti Sekhar Banerjee. 2021. "A Novel Approach to Predict COVID-19 Using Support Vector Machine." In *Data Science for COVID-19*, 351–64. Elsevier. https://doi.org/10.1016/B978-0-12-824536-1.00014-9.
8. Imad, Muhammad, Naveed Khan, Farhat Ullah, Muhammad Abul Hassan, and Adnan Hussain. 2020. "COVID-19 Classification Based on Chest X-Ray Images Using Machine Learning Techniques," 11.
9. Irmak, Emrah. 2021. "COVID-19 Disease Severity Assessment Using CNN Model." *IET Image Processing* 15 (8): 1814–24. https://doi.org/10.1049/ipr2.12153.

10. Jiang, Guoquan, Zhiheng Wang, and Hongmin Liu. 2015. “Automatic Detection of Crop Rows Based on Multi-ROIs.” *Expert Systems with Applications* 42 (5): 2429–41.
11. Khalifa, Nour Eldeen M, Mohamed Hamed N Taha, Aboul Ella Hassanien, and Sally Elghamrawy. 2020. “Detection of Coronavirus (COVID-19) Associated Pneumonia Based on Generative Adversarial Networks and a Fine-Tuned Deep Transfer Learning Model Using Chest X-Ray Dataset.” *ArXiv Preprint ArXiv:2004.01184*.
12. Kim, Tae Jung, Jin Mo Goo, Kyung Won Lee, Chang Min Park, and Hyun Ju Lee. 2009. “Clinical, Pathological and Thin-Section CT Features of Persistent Multiple Ground-Glass Opacity Nodules: Comparison with Solitary Ground-Glass Opacity Nodule.” *Lung Cancer* 64 (2): 171–78.
13. Kondziolka, Doug, William T. Couldwell, and James T. Rutka. 2020. “Introduction. On Pandemics: The Impact of COVID-19 on the Practice of Neurosurgery.” *Journal of Neurosurgery* 133 (1): 1–2. https://doi.org/10.3171/2020.3.JNS201007.
14. Kumar Dubey, Ashutosh, Sushil Narang, Abhishek Kumar, Satya Murthy Sasubilli, and Vicente Garc鞾-D鞾z. 2021. “Performance Estimation of Machine Learning Algorithms in the Factor Analysis of COVID-19 Dataset.” *Computers, Materials & Continua* 66 (2): 1921–36. https://doi.org/10.32604/cmc.2020.012151.
15. Li, Chenxi, Chengxue Zhao, Jingfeng Bao, Bo Tang, Yunfeng Wang, and Bing Gu. 2020. “Laboratory Diagnosis of Coronavirus Disease-2019 (COVID-19).” *Clinica Chimica Acta; International Journal of Clinical Chemistry* 510: 35.
16. Mazzilli, Aldo, Claudio Fiorino, Alessandro Loria, Martina Mori, Pier Giorgio Esposito, Diego Palumbo, Francesco de Cobelli, and Antonella del Vecchio. 2021. “An Automatic Approach for Individual HU-Based Characterization of Lungs in COVID-19 Patients.” *Applied Sciences* 11 (3): 1238. https://doi.org/10.3390/app11031238.
17. McKee, Brady J., Shawn M. Regis, Andrea B. McKee, Sebastian Flacke, and Christoph Wald. 2015. “Performance of ACR Lung-RADS in a Clinical CT Lung Screening Program.” *Journal of the American College of Radiology* 12 (3): 273–76. https://doi.org/10.1016/j.jacr.2014.08.004.
18. Nour, Majid, Zafer Cömert, and Kemal Polat. 2020. “A Novel Medical Diagnosis Model for COVID-19 Infection Detection Based on Deep Features and Bayesian Optimization.” *Applied Soft Computing* 97 (December): 106580. https://doi.org/10.1016/j.asoc.2020.106580.
19. Sun, Nan-Nan, Ya Yang, Ling-Ling Tang, Yi-Ning Dai, Hai-Nv Gao, Hong-Ying Pan, and Bin Ju. 2020. “A Prediction Model Based on Machine Learning for Diagnosing the Early COVID-19 Patients.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.06.03.20120881.
20. Wang, Wenling, Yanli Xu, Ruqin Gao, Roujian Lu, Kai Han, Guizhen Wu, and Wenjie Tan. 2020. “Detection of SARS-CoV-2 in Different Types of Clinical Specimens.” *Jama* 323 (18): 1843–44.

21. Yuan, Mingli, Wen Yin, Zhaowu Tao, Weijun Tan, and Yi Hu. 2020. “Association of Radiologic Findings with Mortality of Patients Infected with 2019 Novel Coronavirus in Wuhan, China.” *PloS One* 15 (3): e0230548.
22. Zebin, Tahmina, and Shahadate Rezvy. 2021. “COVID-19 Detection and Disease Progression Visualization: Deep Learning on Chest X-Rays for Classification and Coarse Localization.” *Applied Intelligence* 51 (2): 1010–21. https://doi.org/10.1007/s10489-020-01867-1.
23. Zhang, Junyong, Yingna Chu, and Na Zhao. 2020. “Supervised Framework for COVID-19 Classification and Lesion Localization from Chest CT.” *Ethiopian Journal of Health Development* 34 (4).
24. Zhao, Jinyu, Yichen Zhang, Xuehai He, and Pengtao Xie. 2020. “Covid-Ct-Dataset: A Ct Scan Dataset about Covid-19.” *ArXiv Preprint ArXiv:2003.13865* 490.
25. Zou, You, Jia-Ni Zou, Ya-Se Zhuang, Bin-Ru Wang, Liu Sun, Shan Xu, Sheng-Lan Li, et al. 2021. “Factors Affecting Recurrent Positive RT-PCR Results in Clinically Cured COVID-19 Patients: A Multicenter Study,” 9.