Baghdad Science Journal

Volume 22 | Issue 2 Article 27

2025

Comparing PCA-Based Machine Learning Algorithms for COVID-19 Classification Using Chest X-ray Images

Hussein Ahmed Ali
Microwave Electronics Research Laboratory, Faculty of Sciences of Tunis, University Tunis El-Manar, Tunis El-Manar, Tunisia AND College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq, hussien.alwaise@uokirkuk.edu.iq

Walid Hariri
College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

Nadia Smaoui Zghal
Labged Laboratory, Department of Computer Science, Badji Mokhtar Annaba University, Annaba, Algeria

Dalenda Ben Aissa
Control and Energy Management Laboratory, (CEM Lab) ENIS, University of Sfax, Sfax, Tunisia

How to Cite this Article

Ali, Hussein Ahmed; Hariri, Walid; Zghal, Nadia Smaoui; and Aissa, Dalenda Ben (2025) "Comparing PCA-Based Machine Learning Algorithms for COVID-19 Classification Using Chest X-ray Images," *Baghdad Science Journal*: Vol. 22: Iss. 2, Article 27.
DOI: https://doi.org/10.21123/bsj.2024.9422

BAGHDAD SCIENCE JOURNAL 2025;22(2):687–705

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

Comparing PCA-Based Machine Learning Algorithms for COVID-19 Classification Using Chest X-ray Images

Hussein Ahmed Ali [1,2,*], **Walid Hariri** [3], **Nadia Smaoui Zghal** [4], **Dalenda Ben Aissa** [1]

[1] Microwave Electronics Research Laboratory, Faculty of Sciences of Tunis, University Tunis El-Manar, Tunis El-Manar, Tunisia
[2] College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq
[3] Labged Laboratory, Department of Computer Science, Badji Mokhtar Annaba University, Annaba, Algeria
[4] Control and Energy Management Laboratory, (CEM Lab) ENIS, University of Sfax, Sfax, Tunisia

ABSTRACT

The rapid spread of the COVID-19 pandemic has strained global healthcare systems, necessitating efficient diagnostic methods. While Polymerase Chain Reaction (PCR) and antigen tests are common, they have limitations in speed and precision. Enhancing the accuracy of imaging techniques, especially Chest X-rays (CXR) and Computerized Tomography (CT) scans, is crucial for detecting COVID-19-related lung abnormalities. CXR, being cost-effective and accessible, is preferred over CT scans, but accurate diagnosis often requires technological support. To address this, an extensive dataset of CXR images categorized into five classes is available on Kaggle. Processing such data involves steps like grayscale conversion, image intensity adjustment, resizing, and feature extraction using Principal Component Analysis (PCA). Machine Learning (ML) techniques, including Decision Tree (DT), Random Forest (RF), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and K-Nearest Neighbors (KNN), are employed for image classification. DT shows the highest accuracy at 88%, outperforming other models like GNB (77%), KNN (71%), SGD (70%), LR (74%), and RF (45%). It consistently excels across assessment metrics such as F1-score, sensitivity, and precision, with an 88% best-weighted average. However, selecting the optimal ML model depends on factors like dataset characteristics and implementation specifics. Thus, careful consideration of these factors is crucial when choosing an ML model for COVID-19 diagnosis via CXR image classification.

Keywords: Chest X-ray (CXR), COVID-19, Decision tree, Gaussian Naïve, Stochastic gradient descent, Bayes, Machine learning

Introduction

Artificial Intelligence (AI) has made significant advancements in medical diagnosis and the development of new medicines.[1,2] AI is projected to significantly impact radiology, providing radiologists with tools for more exact diagnoses and prognoses, ultimately leading to more efficient treatments. Computers, equipped to analyze vast expanses of patient data, are on the verge of replacing radiologists in numerous clinical environments, bringing forth a new era of radiological practice driven by big data and AI. AI has already demonstrated successful applications in treating skin cancer and managing chronic disorders.[3] In the fight against the novel coronavirus, scientists anticipate AI playing a vital role in finding a cure and alleviating the fear associated with the pandemic.[4] The ML offers a robust method for

Received 12 September 2023; revised 19 April 2024; accepted 21 April 2024.
Available online 28 February 2025

* Corresponding author.
E-mail addresses: hussien.alwaise@uokirkuk.edu.iq (H. A. Ali), hariri@labged.net (W. Hariri), nadia.smaoui@isimg.tn (N. S. Zghal), dalenda.benaissa@fst.utm.tn (D. B. Aissa).

https://doi.org/10.21123/bsj.2024.9422

analyzing data in many formats; efficient use necessitates meticulous feature organization. The researcher used this strategy in one study on the online purchases and returns dataset. With a total of 5,659,676 transactions and 15,555 facets, the dataset in issue was quite large.[5]

The COVID-19 pandemic has put tremendous demand on healthcare systems, forcing them to adapt to new techniques. This necessitates the utilization of cutting-edge technologies such as AI to develop intelligent and self-sufficient healthcare solutions.[6,7] COVID-19 stands out among viruses due to its rapid replication and transmission, resulting in a global pandemic within a remarkably short period of time.[8] Extensive research and analysis are ongoing in the medical and healthcare sectors to better understand this rapidly evolving health crisis and develop effective responses.[9] Accurately simulating the spread of COVID-19 remains a critical objective. The gold standard for diagnosis is the detection of viral RNA in sputum by real-time reverse transcription-polymerase chain reaction (RT-PCR) on nasopharyngeal swabs.[10,11] However, these tests can take up to 6 hours to yield results and rely on human intervention while exhibiting a low positive rate in the early stages of illness. Thus, there is a pressing need for rapid and accurate diagnostic methods to bring the pandemic under control as quickly as possible, particularly in the long term when lockdowns are lifted, and widespread testing becomes essential to prevent a resurgence of the virus.[12,13] The prevention of this disease has been approached from various angles. To reduce employee dependency, maintain COVID-19 safety, and cut identity verification expenses, the study developed a COVID-19 Vision system that uses Haar cascades for a real-time face mask detector.[14]

In many countries, COVID-19 testing is primarily available to individuals with disease symptoms. However, it is significant to note that numerous symptomatic patients exhibit more than one sign, making it challenging for national healthcare systems and staff to identify and track potential cases. This burden is particularly overwhelming, even in highly developed nations. To address this crisis, AI algorithms play a crucial role in various aspects of the global health emergency response. AI algorithms are instrumental in the development of drugs and vaccines, as well as in monitoring people's mobility patterns to ensure compliance with social distancing guidelines. These algorithms also assist medical professionals in quickly diagnosing COVID-19 by evaluating CT scans and X-rays of lung conditions, enabling efficient patient tracing.[15,16]

ML algorithms encounter several obstacles while attempting to diagnose COVID-19 using CXR images. These include issues with dataset size, image quality, data augmentation, feature extraction, model selection, and performance evaluation.[17] To overcome these obstacles, it is necessary to enhance the performance of ML algorithms by meticulous preprocessing, feature extraction, model selection, and evaluation procedures. Size, balance or imbalance, and image quality are dataset-related variables that affect how quickly and well ML classification works with CXR images.[18]

The COVID-19 pandemic has urgently needed efficient and accurate disease diagnosis. CXR imaging has emerged as a valuable tool in identifying COVID-19 cases due to its accessibility and cost-effectiveness. To improve the diagnostic procedure, ML applications were used to analyze CXR images and aid in diagnosis and classifying COVID-19 cases. This introduction explores the use of ML in analyzing CXR images for COVID-19 diagnosis, highlighting its potential to improve accuracy, speed up the diagnostic process, and assist healthcare professionals in effectively managing the pandemic. This study examined 14 research articles on COVID-19 and ML, discovering that ML plays a significant function in COVID-19 research, prediction, and discrimination, with supervised learning achieving a testing accuracy of 92.9%, implying its potential inclusion in healthcare programs for assessing and triaging COVID-19 cases. In contrast, recurrent supervised learning may offer even greater accuracy in the future.[19] ML approaches have been extensively employed in the medical arena, particularly in the context of COVID-19, utilizing various imaging systems such as CXR, with applications ranging from diagnosis to forecasting and medication development; however, challenges and limitations still exist, necessitating further research to address issues related to safety and other factors, while Keras remains the most commonly used library in these studies.[20] The critical need for timely and reliable detection of COVID-19 patients was focused on. The study emphasized the advantages of using whole blood count tests for early detection, used ML algorithms for prediction and assessed performance using accuracy, recall, precision, and F-measure metrics.[21,22] ML is a scientific approach that enables computer systems to perform specific tasks without explicit programming, utilizing algorithms and statistical models. ML algorithms are widely applied in various applications, offering the advantage of independent decision-making once trained with data.[23] The study used ML and Deep Learning (DL) algorithms in a multi-test retrospective analytic

approach to detect and assess COVID-19 its progression using CXR features, resulting in a satisfactory "corona score" that demonstrated the high accuracy of advanced AI-based image analysis in diagnosing, quantifying, and monitoring COVID-19.[24]

This article compares ML methods for COVID-19 disease categorization to improve accuracy and implementation time. The main contributions of this paper are summarized as follows:

- Preprocessing: Images are converted to grayscale, density adjusted using Histogram Equalisation and resized for speeding and analysis.
- PCA Feature Extraction: Extracts the most informative features from scaled images.
- Dimensionality Reduction: Images are reduced from two dimensions to one dimension while keeping PCA information to speed up training and reduce hardware requirements.
- Training and Testing ML Models: Prepared CXR images of train and test models.
- Evaluation of Effective ML Classifiers: The study evaluates ML classifiers that can identify COVID-19 cases from five categories using CXR images.

This article is organized as follows: Section two presents related works on COVID-19 categorization from CXR images using ML approaches. Section three is separated into three subsections; each discusses the dataset description, preprocessing techniques, and Feature Extraction Using Principal Component Analysis (PCA). Section 4 covers the ML algorithms used in the study. The fifth section is divided into "Results and Performance of ML Algorithms" and "Evaluation Performance Comparison of ML Algorithms". In the final section, discuss the research's strengths and weaknesses, its practical and theoretical consequences, and plans for the future.

Related works

The capability of ML to manage complicated and enormous datasets is demonstrated here. The global COVID-19 pandemic has highlighted the urgent need for effective detection methods. Several studies have explored the effectiveness of ML methods in achieving high accuracy for COVID-19 diagnosis using CXR images. Recent research has investigated advanced ML approaches, such as KNN, to detect COVID-19 with up to 88% classification rates.[25] To identify COVID-19 cases from CXR images, a comparison of various ML methods, including Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and RFs, was done. The results showed that CNNs outperformed the other algorithms, with the most fantastic accuracy of 95%.[26] Another study used CXR pictures to compare ML methods such as SVM, RF, and gradient boosting for COVID-19 identification. RF scored the highest accuracy of 93%, showing their potential for reliable diagnosis.[27]

CXR radiography can be used to triage non-COVID-19 lung illnesses.[28] However, these investigations frequently require assistance with tiny datasets and the resulting limitations in accuracy. Attempts have been made to construct efficient ML classifiers, with training and testing accuracies in 4-class classification reaching 87%.[29] This survey presents a comprehensive assessment of advanced ML approaches that aid in diagnosing COVID-19 to improve public health.[30] Various datasets containing computed tomography (CXR) scans of healthy subjects, patients with pneumonia, and COVID-19 cases have been used, with mixed results. Two notable approaches were COVID-Net (83% success rate) and a "Naive Bayes" method (87% success rate).[31] KNN has the highest accuracy and weighted average for precision, sensitivity, and F1-score among the ML models tested.[32] The study aimed to develop a Lasso-logistic regression model predicting COVID-19 severity (severe, moderate, and mild), demonstrating 85.9% accuracy and reducing deaths through early detection.[33]

Several other diseases, including Alzheimer's, glaucoma, cancer, and others, have been successfully detected in real-world clinical settings using the same ML methodology that has proven so successful in COVID-19 detection over the last three years.[34] On the other hand, several major obstacles have necessitated the development of more durable devices to train massive datasets effectively using DL algorithms and dealing with the low quality of medical images.[35] Choosing the suitable ML model or DL architecture requires much disease-specific practical experience.

There are some serious limitations to the research that were cited. Some examples are issues like employing ML for CXR image classification without proper CXR image preprocessing methods and imbalanced classes in their datasets. Inconsistent implementation of key techniques led to issues with intensity equalisation, noise removal, resizing, and feature extraction using methods like principal component analysis (PCA). Metrics for evaluation, such as F1-score, recall, accuracy, and precision, require improvement. Furthermore, training ML models usually takes a long period.

These studies demonstrate the significance of ML approaches in achieving high accuracy for COVID-19 diagnosis using CXR images. Using the power of these algorithms, accurate and efficient identification

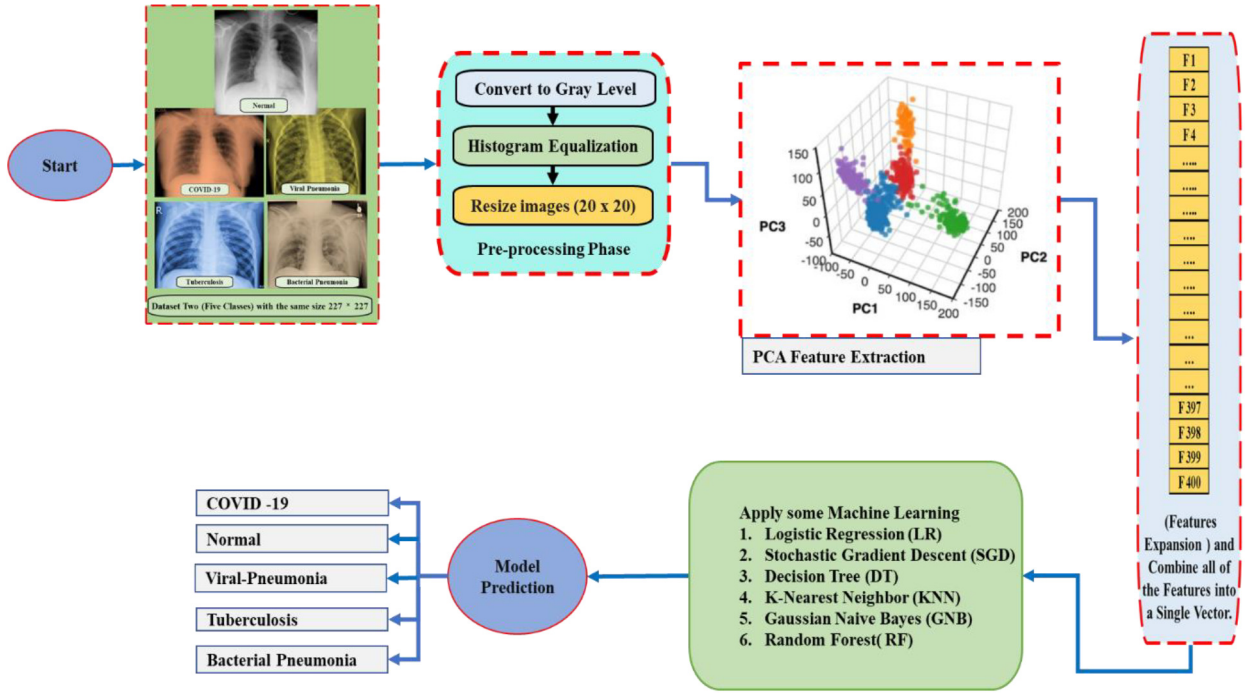

Fig. 1. The methodology diagram illustrates using ML algorithms to classify CXR images.

of COVID-19 instances can be achieved, allowing for timely interventions and reducing disease spread.

Methodology

The COVID-19 epidemic has generated an urgent need to tackle a severe hazard to human health. The correct interpretation and classification of CXR images are critical in diagnosing COVID-19. ML technologies improve imaging tools' capabilities, supporting healthcare professionals' curative efforts. A large number of researchers have classified COVID-19. It takes a fresh approach to ML algorithms to obtain optimal accuracy while requiring little execution time and storage. This section describes the methods used in this study in depth, beginning with the dataset description, CXR imagines preprocessing and feature extraction using PCA, and ending with a description of the ML algorithms used. The methodology provides an in-depth look at the research process. The classification methodology consists of several steps. Initially, preprocessing techniques are applied to convert the image into a grayscale format and adjust its density. Subsequently, the image is resized, and feature extraction is performed using PCA to extract the most informative features. The image is then transformed from two dimensions to one dimension to enhance training speed and minimize hardware requirements. Finally, the prepared images train and test ML models from X-ray images to classify COVID-19. Fig. 1 visually represents the implementation process and taxonomy utilized in our work for dataset preparation before using ML models to classify COVID-19 in CXR images. The remainder of this job is well-organized. Before using ML algorithms, the dataset goes through four preprocessing stages to decrease storage size, with increased execution speed and accuracy of classification. At the outset, the dataset is transformed into grayscale, transitioning all images from three to one channel. The next step in improving the quality of CXR images is to apply Histogram Equalisation to adjust the image intensities. All pictures are shrunk from their original (227 × 227) dimensions to a more manageable (20 × 20) to speed up the execution. The final stage involves utilizing PCA to extract the most relevant characteristics for optimal categorization. After all processes have been performed, the data is split into training (70%) and testing (30%) sets, and several algorithms for ML, such as DT, RF, SGD, LR, GNB, and KNN, are used to classify the data.

Description of the dataset

This study's dataset consisted of five classes derived from three primary datasets. The COVID-19 data were obtained from Cohen et al.'s[36] comprehensive

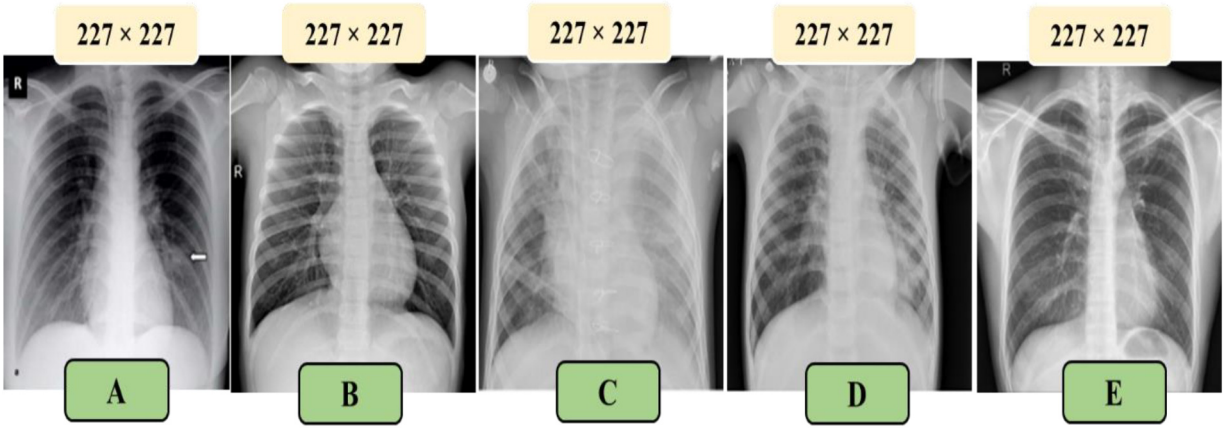

Fig. 2. A selection of CXR images from each dataset class.

X-ray and CT images that included various lung disorders such as COVID-19, SARS, and MEARS. This dataset is regularly updated and comprises 752 X-ray images until June 15, 2020, with the majority (435) depicting cases of COVID-19. Lateral X-rays and CT scans were excluded from this investigation, and incomplete metadata resulted in the omission of gender information for forty-three photographs. On average, the COVID-19 patients in the dataset were approximately fifty-four years old, with approximately two hundred fifty-six males and one hundred thirty-six females. To create balanced sets between pneumonia-positive radiographs (including normal, bacterial, and viral cases) and normal images, the second source utilized a dataset consisting of around 5,863 pictures.[37] Another dataset from the US National Library of Medicine focused on tuberculosis (TB) provided two sets of CXR. This TB dataset supports research on computer-aided diagnosis (CAD) for respiratory diseases, including TB.[36,38] The TB dataset contained a total of 394 images, with 336 originating from China and the remaining 58 sourced from Montgomery County. However, due to the slightly fewer TB X-ray images compared to other classes, the detection performance may be affected by class imbalance.[39] Augmentation techniques were used to resize a random sample of 40 X-ray images to solve this issue. The dataset comprised five classes of CXR images, each with a different number of cases. With a relatively balanced CXR image dataset, this work will classify COVID-19 using multiple ML models. Using a dataset of CXR images that is almost balanced has a positive effect on the accuracy and effectiveness of COVID-19 classification using different ML models, which is a substantial benefit.

- Class 0: Represented 186 verified COVID-19 cases.
- Class 1: Represented 186 Normal cases.
- Class 2: Represented 189 cases of Bacterial Pneumonia.
- Class 3: Represented 173 cases of Viral Pneumonia.
- Class 4: Represented 187 confirmed Tuberculosis cases.

All CXR images maintain a uniform seam size of (227 × 227) for the five classes, ensuring consistency throughout the dataset. With a relatively balanced CXR image dataset, this study will classify COVID-19 using multiple ML models. Using a dataset of CXR images that is almost balanced has a positive effect on the accuracy and effectiveness of COVID-19 classification using different ML models, which is a substantial benefit. Fig. 2 shows a sample CXR image from each dataset class: (A) COVID-19, (B) Normal, (C) Pneumonia-Bacterial, (D) Pneumonia-Viral, and (E) Tuberculosis CXR image.

Preprocessing stage

In ML approaches, one common strategy is to reduce background noise and highlight relevant regions in an image for identification tasks or during the learning phase. This preprocessing method aims to eliminate extraneous data and noisy values. For the model to converge, pixel intensity normalization is performed within the range of [0, 1]. The resized response images are designed to work with the system's architecture and support the ML models. Efficient nets, with low memory and latency costs, are utilized to take advantage of higher-quality response images. This adjustment in response determination

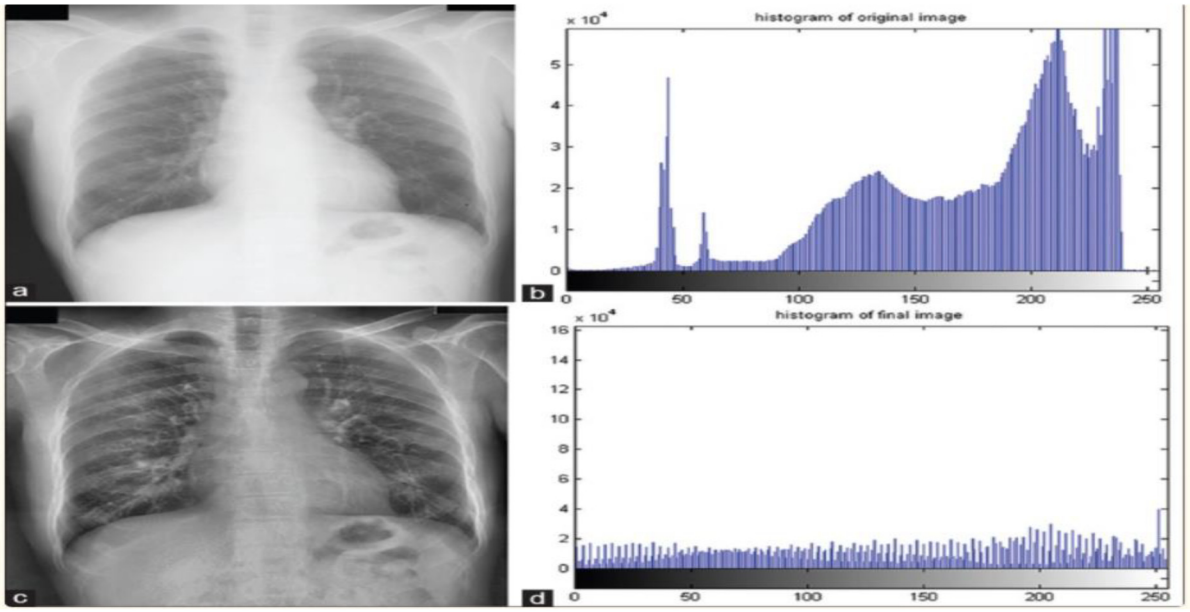

Fig. 3. CXR image following histogram averaging.

can impact the precision of the model. The following steps outline the preprocessing phase of the procedure:

1. Convert the color image (RGB) to a grayscale image. This is achieved using the following Eq. (1):

$$Grayscale\,(i,\ j) = (0.2989 \times r) + (0.5878 \times g) + (0.1140 \times b) \quad (1)$$

Converting the image to grayscale reduces the number of channels from three to one, allowing faster processing than color images.

2. Enhance the image's contrast by applying histogram equalization, as shown in Eq. (2):

$$Hist\,(v) = cut\left(\frac{(cdf_v - cdf_{min})}{(m \times n) - 1}\right) \times (L - 1) \quad (2)$$

The total number of pixels in the image, as m and n, determines the cumulative distribution function (*cdf*). L represents the grey level range, which is 256 levels. Fig. 3 depicts (a) the original chest radiograph image, (b) the original image's histogram, (c) the final improved chest radiograph image, and (d) the final image's histogram.[40]

3. Reduce or reshape the size of the image generated from the previous stages using the following Eq. (3):

$$J_{reshape(x,y)} = \sum_{h=1}^{x}\sum_{w=1}^{y} j\_\,Gray\,(x, y) \quad (3)$$

The original image's width is denoted by w = 227, the height of the original image is denoted by h = 227, and the width and height of the image after resizing are x = 20 and y = 20, respectively, at the same in.[41,42] These preprocessing steps help prepare the image data for further analysis and ML tasks, allowing for improved performance and more efficient processing. Fig. 4 illustrates the various stages of image processing, explicitly focusing on grey-level conversion, histogram analysis, and image resizing to a dimension of (20 × 20) pixels.

Several vital benefits are available during the preprocessing phase of the process. The first step in reducing computing complexity, simplifying data representation, and enhancing image structure and brightness information while eliminating color fluctuations is to convert a color image (RGB) to a grayscale image. Second, histogram equalization makes images seem better by increasing contrast, making details more visible in low-contrast images, and making the dynamic range of pixel intensities more uniform. Finally, there are some advantages to reducing the size of images from 227 × 277 to 20 × 20, including fewer computing demands, less storage space needed, and the possibility of image processing activities being accelerated. In some cases, when the input dimensions

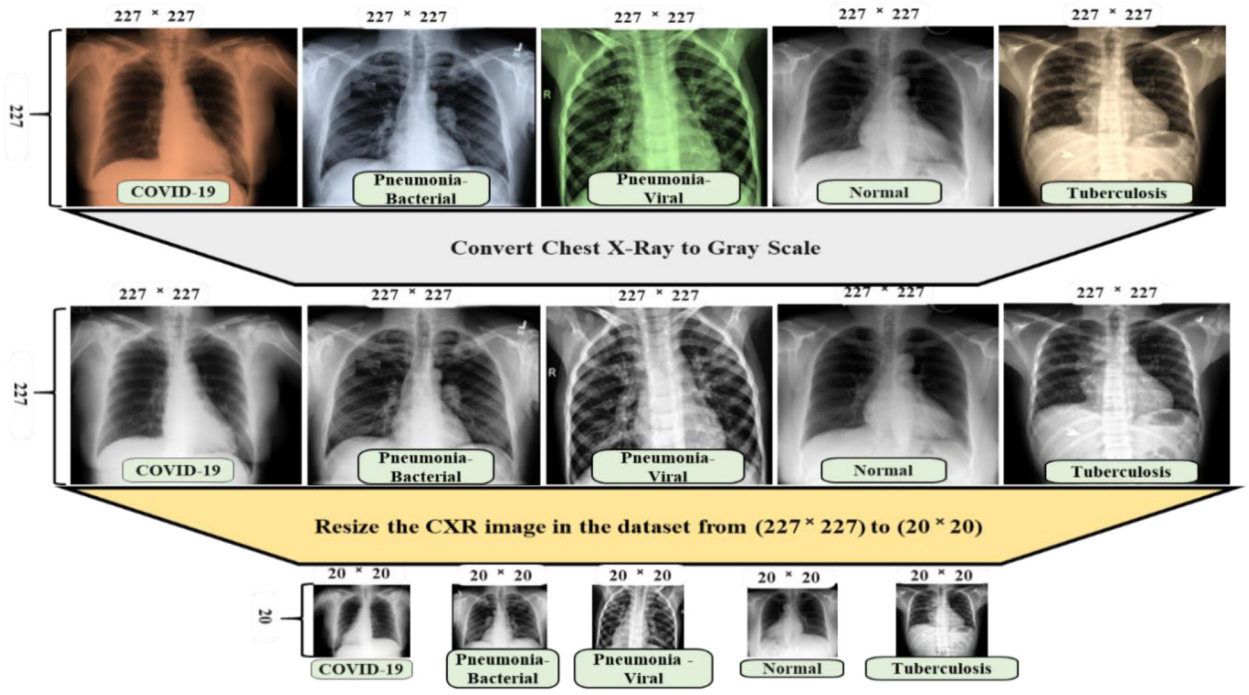

Fig. 4. A sample of CXR images is undergoing three preprocessing stages.

of the analysis or model are smaller than the original image dimensions, this resizing becomes very useful.

Feature extraction

ML models depend on precise feature extraction, which poses significant challenges in accurately diagnosing COVID-19 from X-ray images. If not carefully chosen, inadequate features can result in a suboptimal representation of our data, consequently leading to a less effective classification. The PCA method is selected for feature extraction because of its many advantages. PCA is an invaluable tool for improving the efficiency of ML models used for COVID-19 diagnosis from CXR images.[15] PCA extracts a new collection of principal components from the initial features by lowering the dataset's dimensionality. The amount of information to preserve is significantly affected by the selection of principal components, and the first principal component captures the most variance compared to others. Choosing the correct number of significant components allows for the optimal retention of relevant information. ML models trained with CXR images have significantly improved their ability to diagnose COVID-19.

PCA is a commonly used statistical technique for extracting features and image representation. PCA aims to minimize the dimensionality of high-dimensional data while maintaining as much original information as a CXR image. After preprocessing a set of images, PCA is then used to process another set of images so that its feature extraction knowledge can be used.[43] PCA is a feature extraction and dimensionality reduction technique by utilizing an orthogonal transformation to convert potentially correlated observations into linearly uncorrelated variables. It is a practical method for feature extraction in pattern recognition.[44] Fig. 5 shows the case of the PCA approaches. The main objective of PC is to represent patterns with a reduced number of features, reducing dimensionality while retaining crucial discriminative information. PCA is a traditional pattern recognition approach for feature extraction and data representation. Its purpose is to capture the essence of patterns with fewer features and reduce the dimensionality of the feature space while retaining critical discriminative information. One notable application of PCA is Eigenface, which utilizes PCA techniques to extract characteristic features from facial images. It represents a given face as a linear combination of "eigenfaces" obtained through feature extraction.[45] The PCA is a linear modification that can decrease the number of dimensions in a dataset. Maximizing the data's variance helps achieve this goal, producing a vector of orthogonal basis groups with no correlations.[46,47]

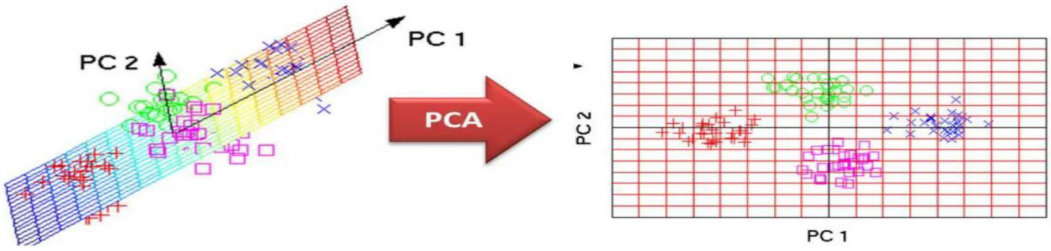

Fig. 5. A comprehensive primer on principal component analysis.

Taking into account d data points, where n is the number of dimensions in the dataset, and the fact that supplied by $z_1,\ z_2,\ z_k \in R^n$, PCA is carried with using the following techniques:

- The m-dimension mean vector *MV* may be computed as Eq. (4): To get the m-dimensional mean vector MV, the average all data points in the dataset, z_1, z_2, . . . , z_k. The algorithm determines the mean by adding all the data points and dividing by the total number of data points (d).

$$MV = \frac{1}{d}\sum_{i=1}^{d} z_i \tag{4}$$

- The covariance matrix CM for the observations is Eq. (5): The dataset's relationships are represented in the covariance matrix CM. For its calculation, to take the dot product of the vectors that arise from subtracting the mean vector (MV) from each data point (z_i). Can may find their covariance matrix for each set of data points by adding up their outer products and dividing by (d).

$$CM = \frac{1}{d}\sum_{i=1}^{d} (z_i - MV)\, x_i - MV)^d \tag{5}$$

- The eigenvalues and eigenvectors are calculated based on CM.

Covariance matrices (CMs) have their eigenvalues and eigenvectors determined. Each principal component's eigenvalue and eigenvector indicate the variation it explains and the direction in which each principal component is located, respectively.

- This is performed through the PCA method, which involves linear transformations to reduce data dimensionality. Eq. (6): Everyone uses the PCA technique to reduce dimensionality. By adding together the original data points (z_1, z_2, . . . , z_d) and multiplying them with the associated coefficients (a_d1, a_d2, . . . , a_dd), the converted data point (y_d) in the reduced-dimensional space can be obtained.

$$y_d = a_{d1}z_1 + a_{d2}z_2 + \cdots + a_{dd}z_d \tag{6}$$

Machine learning algorithms

CXR is a type of medical imaging that is crucial in the global fight against COVID-19. Recent advancements in ML technologies have significantly enhanced CXR imaging capabilities and have proven valuable tools for medical professionals. This study used ML models to identify COVID-19 instances, Pneumonia-Bacterial, Pneumonia-Viral, and normal occurrences in CXR images. The preprocessing steps include adjusting grey levels, histogram equalization, and resizing, followed by feature extraction on the dataset using PCA. Upon completion of the preprocessing steps and the application of PCA for feature extraction, resulting in 400 features for each CXR image, the dataset is subsequently separated into 70% for training and 30% for testing purposes. They utilized several ML algorithms, including DT, RF, SGD, LR, GNB, and KNN, to evaluate their performance and effectiveness within different prediction models. The results of these algorithms were computed and evaluated, and to provide further insight, a comparison was made with recently published COVID-19 detection models. The application of these algorithms in our experiments is depicted in Fig. 6.

Decision tree DT

DT classifiers are widely recognized as one of the prominent approaches for data classification, serving as effective representation models.[48] The DT comprises core nodes that function as data pattern tests and leaf nodes that serve as data pattern categories. These tests are run across the tree to get the best output for a given input pattern. DT algorithms find applications in various domains.[49] This supervised ML algorithm is capable of solving classification and regression problems. It is known for its simplicity

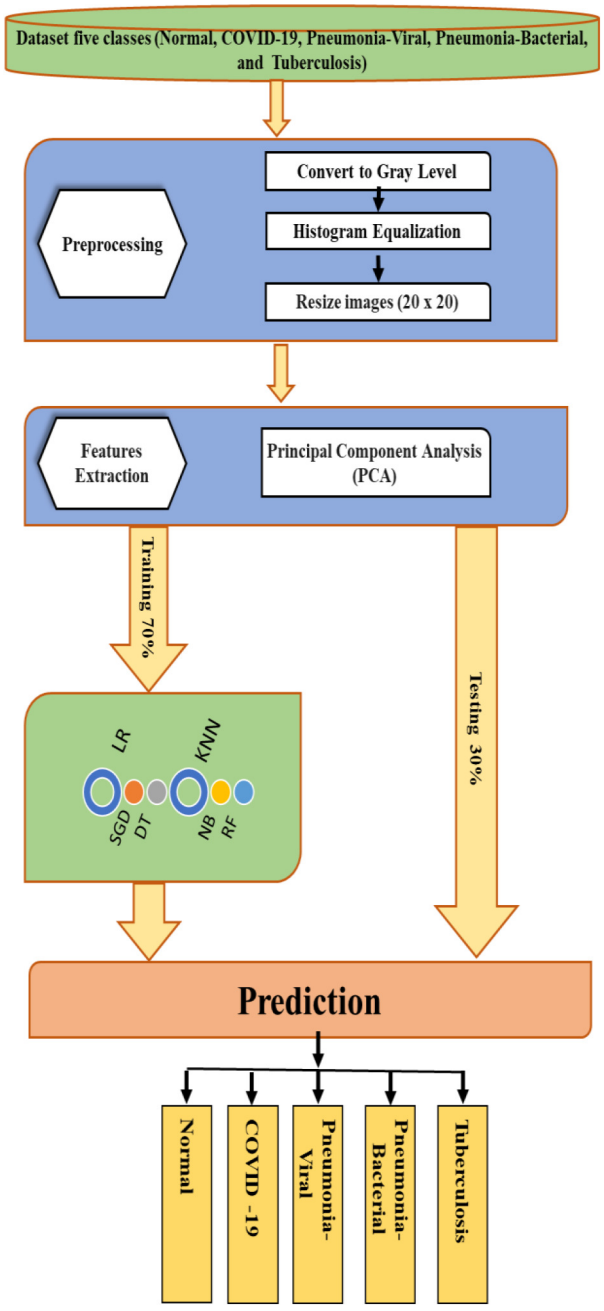

Fig. 6. The flowchart of the ML algorithms to classify CXR images.

and effectiveness in classification tasks. Mathematically, the DT algorithm involves understanding the concept of entropy (H) before delving into the calculation of Information Gain (IG), as depicted in Eq. (7):

$$Entropy\,(H) = \sum_{i=1}^{c} - p_i * log_2 \tag{7}$$

- H represents entropy.
- c is the number of classes or categories.
- p_i is the probability of occurrence of the i-th class. p_i)
- log_2 is the base-2 logarith2.

The entropy and Information Gain (IG) formulas and their parameter values. Entropy is essential in the decision-making process of a DT since it determines data segmentation and boundary construction. It is used to assess the impureness or randomness of a dataset. Conversely, Information Gain (IG) is used to determine the best feature for splitting at each stage of tree development. The capacity of DT classifiers to effectively handle randomness in performance outcomes is well known. The formula for IG is known as Eq. (8).

$$IG\,(Y/X) = H\,(Y) - H\,(Y/X) \tag{8}$$

- $IG\ (Y \mid X)$ represents the IG when you split the dataset Y based on the attribute X.
- $H\ (Y)$ is the entropy of the original dataset.
- $H\ (Y \mid X)$ is the conditional entropy of Y given X, which is the entropy of Y after the dataset has been split based on the attribute X.

Random forest (RF)

The most common application of RF as a supervised ML approach enables the solution of classification and regression problems.[50] RF constructs decision trees from diverse samples and employs majority voting for classification or averaging for regression. It is widely recognized as an efficient classification method and has been successfully applied for COVID-19 prediction in numerous studies.[8] This study employs an ML model, specifically RF, to identify critical features for distinguishing COVID-19 cases from non-COVID-19 cases. When performing RF on classification data, it is important to consider the Gini index, which determines the connections between nodes in a DT branch. The Gini index, as represented by Eq. (9), computes the Gini impurity of each branch based on class distribution and probabilities, thereby aiding in determining the more likely branch outcome.

$$Gini = \sum_{i=1}^{c} -(p_i)^2 \tag{9}$$

c is class count.

p_i *is the probability of class i in the dataset*.

Calculates the Gini impurity for each branch based on the class and its probability, helping determine the likelihood of occurrence for each branch.

Stochastic gradient descent (SGD)

SGD is a common algorithm in many ML approaches, particularly as the foundation for neural networks.[51] SGD is an iterative procedure that begins at an arbitrary idea on a function and steadily descends its slope until it reaches the minimum point. In the case of SGD, the parameters are given a random beginning value, and partial derivatives concerning each feature are computed.[51] SGD excels in proper convex loss functions using linear classifiers and regressors.[52] As a result, ML classifiers, notably SGD adaptive classifiers, are used to examine data with appropriate tools. Our work used PCA to extract key features and achieve maximum accuracy while utilizing SGD to diagnose COVID-19. The SGD formula is as follows:

$$two\ (b) = \frac{1}{2n}\sum_{i=1}^{n}\left(f(xi) - Yi\right)^2 \tag{10}$$

The learning rate (η) is typically chosen as 0.1 or 0.01. The new parameters are updated using the following Eq. (11):

$$New\ params\ = Old\ params - \eta * Derivative\ J\ (b) \tag{11}$$

Logistic regression (LR)

LR is still one of the most popular ML techniques, especially for binary classification jobs. LR calculates the likelihood of a specific result based on the input factors. The cost function determines the optimal values for 0 and 1 to construct the best-fit line for the data points provided.[53] The cost function evaluates the model's performance in linear regression by optimizing the regression coefficients or weights. The Mean Squared Error (MSE) cost function computes the usual squared error between the expected and actual data values.[54] It may be deduced that ML classification algorithm models, such as LR, can be used to predict COVID-19 patients.[55] LR uses the sigmoid function to convert predicted values into probabilities. The sigmoid function transforms any actual value into a rate between "0 and 1". The sigmoid function formula is as follows 12:

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{12}$$

- $f(x)$ is the output value between 0 and 1.
- e is the base of the natural logarithm (approximately equal to 2.71828).
- x is the input value.

The cost function is MSE, which measures the average squared difference between anticipated and actual values. The equation represents the Mean Squared Error (MSE) formula 13:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i\ - y_i\ )^2 \tag{13}$$

- n is the number of data points.
- x_i represents the actual (observed) value for the i-th data point.
- y_i represents the predicted value for the i-th data point.

K-Nearest neighbors (KNN)

The current slow learning strategy is KNN, based on the traditional KNN algorithm.[56] The KNN classifier, as previously stated, is a frequent version of the closest neighbor technique that involves categorizing an unknown sample based on the votes of k nearest neighbors rather than simply one nearest neighbor. KNN is a supervised ML algorithm.[55] That describes in full the stages involved in the KNN algorithm. To forecast COVID-19 patients, can also use the KNN method, one of the ML classification models. The KNN method employs the following Euclidean distance formula:

$$the\ d\ (p, q) = (q, p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i\ )^2} \tag{14}$$

Gaussian naive bayes (GNB)

GNB is one of the most straightforward categorization algorithms,[57] and Naive Bayes (NB) classifiers rely on Bayes' Theorem. These classifiers adopt a strong assumption of feature independence, treating the worth of one feature as distinct from the worth of any other feature. NB classifiers are effectively taught in a supervised learning framework and are simple to create and deploy, making them useful in various real-world scenarios. When working with continuous data, it is expected to assume that the values for each class are regularly distributed (Gaussian).[58] To classify CXR images, this study employs the GNB algorithm specifically created for COVID-19 identification. GNB models make use of continuous values with Gaussian (normal) distributions. When working with continuous data, assuming that the values associated with each class follow a normal distribution is expected. The feature likelihood estimation can be

represented as follows:

$$p(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (15)$$

- ➢ $p(x_i \mid y)$ This represents the likelihood of the variable x_i given y, under certain conditions.
- ➢ $\frac{1}{\sqrt{2\pi\sigma_y^2}}$ Normalisation ensures the entire probability integrates to 1. It is the reciprocal of the standard deviation times the square root of 2π.
- ➢ $exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$ The exponential term is the Gaussian distribution likelihood term. It shows the probability of a given x_i given y.

Evaluations metrics

A confusion matrix is a critical tool for assessing the performance of ML algorithms, particularly in classification tasks. It presents a complete overview of the algorithm's predictions about the actual labels of the data. This matrix is made up of four main components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which enable the computation of critical metrics such as accuracy, precision, recall, and F1 score.[59] The confusion matrix aids in understanding the strengths and weaknesses of a model, enabling researchers to make informed decisions for improving its performance. At this stage, the focus is on determining the accuracy of the classifier. Evaluating the classifier's effectiveness involves assessing how well the anticipated class labels align with the observed ones.

- **Accuracy**: Also known as model viability, it indicates the suits of correct forecasts to total predictions:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (16)$$

- **Precision**: It measures how accurately the classifier assigns documents to a specific category. Class precision is quantified as:

$$Precision = \frac{TP}{(TP + FP)} \quad (17)$$

- **Recall**: It indicates the classifier's ability to identify documents belonging to a particular class correctly. Class recall can be calculated as:

$$Recall = \frac{TP}{(TP + FN)} \quad (18)$$

- **F1-Score**: This metric evaluates the balance between precision and recall. A high F1 score indicates satisfactory overall performance of the system. It is computed as:

$$F1\text{-}Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \quad (19)$$

Results and discussion

Results and performance of ML algorithms

After completing the preprocessing stages and using PCA for feature extraction, resulting in 400 features per CXR image, 70% of the dataset was set aside for training and 30% for testing. Following that, multiple ML techniques, such as DT, RF, SGD, LR, GNB, and KNN, were used to diagnose COVID-19 using the CXR images. The outcomes of these algorithms were assessed, and their performance was evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. In the following, the present the results obtained using each ML technique, emphasizing the parameters associated with each one.

Decision tree (DT)

Several parameters are critical for deciding the model's performance and interpretability in COVID-19 diagnosis using CXR images. DTs are an ML model that uses data-learned rules to generate decisions.

- Criterion: A split's quality can be evaluated using entropy Eq. (7).
- Maximum depth of the tree is 10
- Sets minimum leaf samples to 2; node split requires 1 sample.
- Information Gain (IG): Employed by Eq. (8).

Table 1 details the effectiveness and reliability of the DT algorithm for COVID-19 CXR image diagnosis. It provides information about the model's capacity to diagnose COVID-19 occurrences and aids in calculating critical metrics, including accuracy, precision, recall, and F1-score.

Random forest (RF)

Important parameters in RF approach are used to find features that can distinguish COVID-19 situations from those that do not:

- The number of trees in the forest is 100.
- A criterion for splitting: 70% for training and 30% for testing.
- The maximum depth of each tree is 10.
- Minimum samples are required to split a node 5.

Table 1. DT evaluation metrics for COVID-19 diagnosis using CXR images.

Algorithm	Accuracy	Class	Precision	Recall	F1-score
Decision Tree (DT)	0.88	1	0.96	0.94	0.95
		2	0.89	0.95	0.92
		3	0.83	0.79	0.81
		4	0.79	0.82	0.81
		5	0.95	0.92	0.93
weighted average			0.89	0.88	0.88

Table 2. Evaluation metrics of RF for detection of COVID-19 using CXR image.

Model ML	Accuracy	Class	Precision	Recall	F1-score
Random Forest (RF)	0.45	1	1.00	0.31	0.48
		2	0.00	0.00	0.00
		3	0.13	0.22	0.16
		4	0.17	0.69	0.27
		5	1.00	0.83	0.90
weighted average			0.86	0.45	0.54

Table 3. Evaluation metrics of SGD for diagnosis of COVID-19 using CXR images.

Model ML	Accuracy	Class	Precision	Recall	F1-score
Stochastic Gradient Descent (SGD)	0.70	1	0.80	0.91	0.85
		2	0.76	0.94	0.84
		3	0.00	0.00	0.00
		4	0.96	0.49	0.65
		5	0.93	0.70	0.80
weighted average			0.89	0.70	0.76

The relationships between nodes in a DT branch are heavily influenced by these parameters and the Gini index, as shown in Eq. (9). To help forecast the more likely outcome of a given branch, the Gini index makes it easier to compute the Gini impurity using class distribution and probabilities. Table 2 deduces some information from the specified metrics. Each class's precision, recall, and F1-score values demonstrate how well the system recognizes COVID-19 occurrences in various classes. The accuracy shows the RF algorithm's correctness in diagnosing COVID-19 CXR images.

Stochastic gradient descent (SGD)

SGD is an essential technique for neural networks and a cornerstone of many ML methods; crucial parameters and features of COVID-19 in this domain include:

- Learning Rate (η): This option controls the step size set to 0.01 during iterative descent.
- Number of Iterations (Epochs): 1000; SGD is defined as a process that iteratively approaches the minimal point by gradually reducing the slope of a function.

These parameters help SGD work, especially with suitable convex loss functions, linear classifiers, and regressors. In this work, PCA extracts crucial features to maximise COVID-19 diagnosis accuracy with SGD. Eq. (10), which represents SGD, describes the iterative optimisation process, whereas Eq. (11) describes the parameter update method using the chosen learning rate (0.01).

Table 3 presents the SGD algorithm's effectiveness and reliability for COVID-19 CXR image diagnosis, offering a more comprehensive understanding of its performance in successfully recognizing COVID-19 instances.

Logistic regression (LR)

In LR, a model that calculates the likelihood of a certain result depending on input factors, relevant parameters and aspects include:

- Regularization Term (Set C Parameter:(1.0): The C parameter affects regularisation strength and penalises big coefficients to prevent overfitting.
- Solver ('liblinear'): This solver algorithm, 'liblinear,' affects the LR model optimisation procedure.
- Maximum Iterations: set (100) iterations or epochs controls optimisation algorithm convergence during model training.
- Sigmoid Function: The sigmoid function Eq. (12) converts projected values into probabilities between 0 and 1.

Table 4. Evaluation metrics of LR for detection of COVID-19 using CXR images.

Model ML	Accuracy	Class	Precision	Recall	F1-score
Linear Regression (LR)	0.74	1	0.80	0.91	0.85
		2	0.79	0.89	0.84
		3	0.85	0.55	0.67
		4	0.32	0.74	0.45
		5	0.89	0.72	0.80
weighted average			0.80	0.74	0.75

Table 5. Evaluation metrics of KNN for detection of COVID-19 using CXR images.

Model ML	Accuracy	Class	Precision	Recall	F1-score
K-Nearest Neighbors (KNN)	0.71	1	0.76	0.87	0.81
		2	0.83	0.88	0.85
		3	0.62	0.55	0.58
		4	0.51	0.60	0.55
		5	0.82	0.69	0.75
weighted average			0.80	0.74	0.72

- Mean Squared Error (MSE): MSE cost function Eq. (13) measures model performance by calculating the average difference between expected and actual values.

As predicted, these parameters help the LR model predict outcomes, including COVID-19 patients. The regularization term, solver choice, and maximum iterations affect optimization, while the sigmoid function and MSE transform predictions and evaluate model correctness.

The LR algorithm assessment metrics for the analysis of COVID-19 using the CXR image are in Table 4. These metrics are critical in measuring the effectiveness and reliability of the LR algorithm for COVID-19 CXR image diagnosis, allowing for a thorough assessment of its performance in accurately detecting COVID-19 cases.

K-Nearest neighbor (KNN)

Important KNN algorithm parameters for assessing its performance and dependability in COVID-19 CXR image diagnosis include:

- Number of Neighbors (k): 'k' is set to 5 to reflect the nearest neighbors used for forecasts. This setting affects decision boundary granularity and model sensitivity to local patterns.
- Distance Metric (Euclidean): The Euclidean distance metric Eq. (14) is used. This metric measure instance similarity by calculating the straight-line distance between data points in multidimensional space.
- Weight Function (Distance): is 'distance,' indicating that closer neighbours impact prediction more than faraway ones. This weighting approach improves algorithm performance by prioritising neighboring instances during decision-making.

These parameters influence the KNN algorithm's learning approach and prediction power for COVID-19 diagnosis using CXR pictures. The supervised ML algorithm KNN uses these criteria, mainly the number of neighbors and the distance measure, to categorise unknown samples based on their k nearest neighbors' votes. The parameters help the system adapt to medical picture data.

The KNN algorithm assessment metrics for COVID-19 CXR image diagnosis are in Table 5. These criteria are critical in measuring the effectiveness and reliability of the KNN algorithm for diagnosis of COVID-19 using CXR images, allowing for a thorough assessment of its performance in accurately detecting COVID-19 instances.

Gaussian naive bayes (GNB)

Important features and parameters in the field of GNB for COVID-19 detection using CXR images are summarised below:

- Feature Independence Assumption in COVID-19 Like Naive Bayes (NB) classifiers, GNB assumes feature independence and treats each feature's value as distinct. GNB is effective and simple because this assumption simplifies modelling.
- Supervised Learning Framework: Supervised learning helps apply GNB to real-world situations. Supervised learning trains the model using labelled data to predict new occurrences.
- Utilization of Gaussian Distributions: When working with continuous data, GNB assumes a Gaussian (normal) distribution for class values. For COVID-19 CXR image classification, this trait is especially useful for continuous features.
- Feature Likelihood Estimation: Eq. (15) shows how GNB estimates feature likelihood using the

Table 6. Evaluation metrics of GNB for detection of COVID-19 using CXR images.

Model ML	Accuracy	Class	Precision	Recall	F1-score
Gaussian Naive Bayes (GNB)	0.77	1	0.96	0.96	0.96
		2	0.87	1.00	0.93
		3	0.04	0.33	0.07
		4	0.92	0.48	0.63
		5	1.00	0.90	0.95
weighted average			0.92	0.77	0.81

Gaussian distribution for continuous values. The probability density function for continuous data is calculated using the mean in this equation. Incorporating the mean (μy) and standard deviation (σy) parameters for each class (y).

The parameters of GNB describe its technique, which focuses on assuming feature independence, being suitable for supervised learning scenarios, and using Gaussian distributions to handle continuous data for COVID-19 identification in CXR images.

Table 6 provides a more comprehensive evaluation of the GNB algorithm's performance in COVID-19 diagnosis utilizing CXR images. This table provides useful information about the model's capacity to recognize COVID-19 occurrences and aids in calculating critical assessment metrics such as accuracy, precision, recall, and F1-score.

Evaluation performance comparison of machine learning algorithms

From the above results, it can be observed that the DT algorithm has the highest accuracy (0.88) and F1-score (0.88) among the evaluated algorithms. It also demonstrates high precision and recall for Class 1, indicating good performance in correctly identifying COVID-19 cases. The LR and SGD algorithms also show competitive results with relatively high accuracy and F1-scores. On the other hand, KNN and RF algorithms exhibit lower accuracy and F1-scores than the different algorithms. RF has deficient performance, especially regarding precision and recall for Class 1. It's important to note that the quality and amount of the dataset, feature extraction approaches, hyperparameter adjustment, and the unique properties of the COVID-19 CXR pictures utilized can all impact the success of these algorithms. Further tuning and experimentation may be required to increase the algorithms' accuracy in diagnosing COVID-19. Fig. 7 is most likely a comparison of the performance of these algorithms based on specified evaluation metrics. Of course, creating a new DL model for COVID-19 diagnosis using similar tools can still be improved. Using CXR pictures to develop a DL for diagnosing COVID-19 is a potential strategy for improving speed and accuracy.

The outcome analysis shows various ML models exhibit differing performance levels. Using balanced criteria for precision, recall, and F1-score, RF showed a respectable accuracy of 45%. With a 70% improvement in accuracy, SGD demonstrated an admirable harmony between recall, precision, and F1-score. With identical precision, recall, and F1-score values, LR and KNN attained 74% and 71% accuracy, respectively. GNB's 77% accuracy demonstrated its performance compared to other models, showcasing its excellent recall, F1-score, and precision. Remarkably, the DT stood out as the best-performing model, surpassing all others with an impressive accuracy of 88% and demonstrating its efficacy in the specific context through solid precision, recall, and F1-score metrics. Using DT to classify CXR images has several advantages. One is that it provides a technique for image classification that is transparent and easy to understand. Regarding CXR data, DT is the way to manage numerical and categorical characteristics. This allows for good image category categorization. Their ability to detect non-linear connections in the image features improves the precision of classification. Furthermore, DT handles missing values in CXR datasets with aplomb. They are useful in classifying CXR images because they are versatile, easy to deploy, and can discover essential elements. Table 7 compares prior research that compared the use of CXR images for COVID-19 identification.

ML model results suggest intriguing research avenues. Adjusting CXR image quality, dataset sizes, and preprocessing approaches may improve COVID-19 diagnosis accuracy. These research directions aim to improve COVID-19 diagnostics using ML, considering image quality, dataset features, and preprocessing methods. In addition to Traditional PCA, investigating other CXR image feature extraction methods are crucial to improving diagnostic procedures. While Traditional PCA was used for dimensionality reduction in this work, it is essential to note that the choice of PCA variations, such as sparse, kernel, Incremental PCA (IPCA), or Robust PCA, can affect the principal component properties. Despite these changes, PCA transforms data into orthogonal

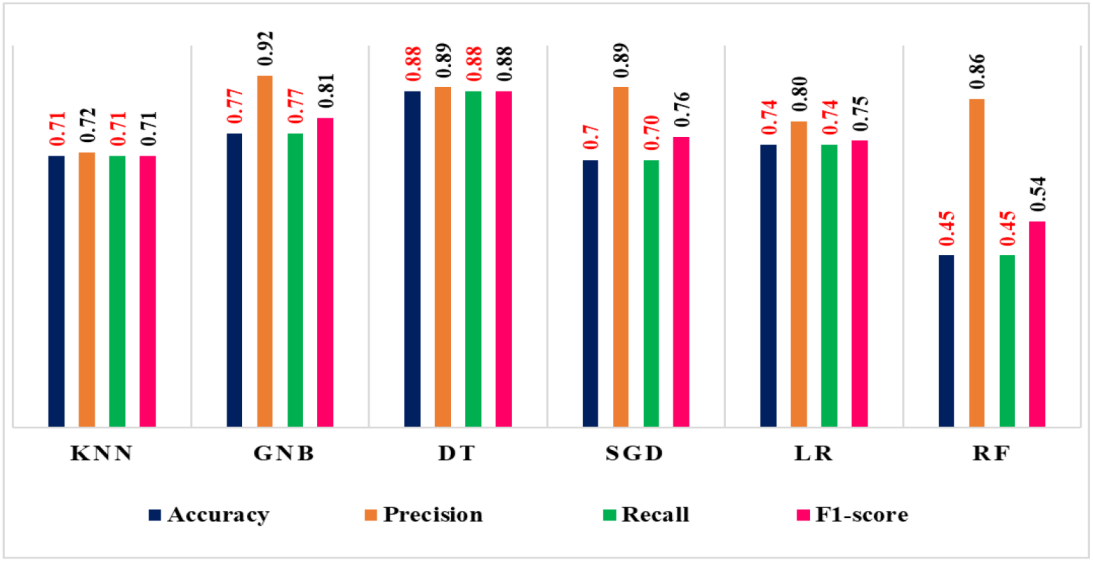

Fig. 7. Conclusion of the evaluation of ML algorithms for accuracy, precision, recall, and F1-score.

Table 7. Comparing previous works using chest X-Ray images for COVID-19 detection.

References	No. of class and N0 of chest X-rays	Technique	Accuracy (%)
22	3-Class/8851 CXR/all images COVID-19 (498)	Machine learning classifiers	76
29	4-class Normal (1341), pneumonia (1345), COVID-19 (757) and Bacterial pneumonia (2782)	Multi-stage framework	87
31	3-Class/1345 CXR/Normal (109), pneumonia (1126), and COVID-19 (110)	Naïve Bayes	87
31	4-class/not available.	COVID-Net	83
33	2-Class-Normal, and COVID-19	Lasso-Logistic regression	85

vectors (principal components) in feature space. In future work, can intend to extend our proposed method using deep-feature extractors using re-trained models such as Google-Net, ResNet, Xception and another model.[60,61] These models need larger datasets to extract the best features and to ensure an accurate classification using our proposed ML techniques. Furthermore, Vision Transformers can be applied to address the limitations of CNNs, as proposed in.[61]

Conclusion

The CXR images can aid in detecting COVID-19 -related diseases, although their significance is typically overlooked. Using chest CXR images, this study examined multiple ML algorithms for accurate COVID-19 diagnosis. Histogram equalization was employed to enhance the CXR images, followed by resizing images and feature extraction utilizing PCA techniques. PCA's main advantage is its ability to reduce large dataset's dimensionality while maintaining the most crucial variance information. Various ML models were employed after completing all preprocessing steps on CXR images and identifying optimal features. DT has the most excellent weighted average for all parameters among the six classification algorithms tested (DT, RF, SGD, LR, GNB, and KNN), showing higher performance than the other models. The DT stood out as the best-performing model, surpassing all others with an impressive accuracy of 88% and demonstrating its efficacy in the specific context through solid precision, recall, and F1-score metrics.

Since our dataset is almost balanced, in the realm of creating an automated system for classifying medical images, addressing imbalanced data poses a notable hurdle. This challenge emerges when there's a substantial discrepancy in sample numbers among various classes, impacting the model's accuracy by favoring the majority class over accurately categorizing the minority class. Additionally, optimizing preprocessing methods with recently updated PCA extraction features should enhance accuracy in classifying COVID-19 in glossy CXR images. To improve the performance of CXR imagine classification, future work will involve including deep-feature extractors, re-trained models (such as Google Net, ResNet, Xception, etc.), increasing datasets, and applying specific transfer learning approaches. Using similar techniques, future research

can investigate DL models for COVID-19 diagnosis. Alternate and updated PCA versions or other datasets could be investigated.

Acknowledgement

The researchers would like to offer their heartfelt gratitude and appreciation to everyone who helped and contributed to the successful completion of this study.

Author's declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Authors sign on ethical consideration's approval.
- No animal studies are present in the manuscript.
- The project was approved by the local ethical committee at the University of Kirkuk.

Author's contribution statement

H.A.A. conceived the idea for the article and provided overall experiments under supervision of W.H, N.S and D.B.A. W.H. meticulously monitored the progress, offering feedback to enhance scientific and linguistic aspects. W. H, N.S.Z. and D.B.A. conducted the revision and proofreading of the article. All authors discussed the results and contributed to the final manuscript.

References

1. Kaheel H, Hussein A, Chehab A. AI-Based image processing for COVID-19 detection in chest CT scan images. Front Comms Net. 2021;2(Aug):1–12. https://doi.org/10.3389/frcmn.2021.645040.
2. Too J, Mirjalili S. A hyper learning binary dragonfly algorithm for feature selection: A COVID-19 case study. Knowl. Based Syst. 2021;212:106553. https://doi.org/10.1016/j.knosys.2020.106553.
3. Shen C, Yu N, Cai S, Zhou J, Sheng J, Liu K, *et al.* Quantitative computed tomography analysis for stratifying the severity of Coronavirus Disease 2019. J Pharm Anal. 2020;10(2):123–9. https://doi.org/10.1016/j.jpha.2020.03.004.
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. https://doi.org/10.1038/nature21056.
5. Soni M, Shnan MA. Scalable neural network algorithms for high dimensional data. Mesopotamian J Big Data. 2023;1–11. https://doi.org/10.58496/MJBD/2023/001.
6. Adadi A, Lahmer M, Nasiri S. Artificial intelligence and COVID-19: A systematic umbrella review and roads ahead. J King Saud Univ Inf Sci. 2022;34(8):5898–920. https://doi.org/10.1016/j.jksuci.2021.07.010.
7. Bachtiger P, Peters NS, Walsh SLF. Machine learning for COVID-19—asking the right questions. Lancet Digit Heal. 2020;2(8):e391–2. https://doi.org/10.1016/s2589-7500(20)30162-x.
8. Gupta VK, Gupta A, Kumar D, Sardana A. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. Big Data Min Anal. 2021;4(2):116–23. https://doi.org/10.26599/bdma.2020.9020016.
9. Shaikh F, Andersen MB, Sohail MR, Mulero F, Awan O, Dupont-Roettger D, *et al.* Current landscape of imaging and the potential role for artificial intelligence in the management of COVID-19. Curr Probl Diagn Radiol. 2021;50(3):430–5. https://doi.org/10.1067/j.cpradiol.2020.06.009.
10. Elmokadem AH, Mounir AM, Ramadan ZA, Elsedeiq M, Saleh GA. Comparison of chest CT severity scoring systems for COVID-19. Eur Radiol. 2022;1–12. https://doi.org/10.1007/s00330-021-08432-5.
11. Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, *et al.* COVID-19 patient health prediction using boosted random forest algorithm. Front public Heal. 2020;8:357. https://doi.org/10.46632/daai/3/2/13.
12. Abbood EA, Al-Assadi TA. GLCMs based multi-inputs 1D CNN deep learning neural network for COVID-19 texture feature extraction and classification. Karbala Int J Mod Sci. 2022;8(1):28–39. https://doi.org/10.33640/2405-609x.3201.
13. Ufuk F, Demirci M, Uğurlu E, Çetin N, Yiğit N, SarıT. Evaluation of disease severity with quantitative chest CT in COVID-19 patients. Diagn Interv Radiol. 2021;27(2):164. https://doi.org/10.5152/dir.2020.20281.
14. Kareem OS. Face mask detection using haar cascades classifier to reduce the risk of Coved-19. Int J Math Stat Comput Sci. 2024;2:19–27. https://doi.org/10.59543/ijmscs.v2i.7845.
15. Rasheed J, Hameed AA, Djeddi C, Jamil A, Al-Turjman F. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdiscip Sci Comput Life Sci. 2021 Mar 1;13(1):103–17. https://doi.org/10.1007/s12539-020-00403-6.
16. Gouda W, Almurafeh M, Humayun M, Jhanjhi NZ. Detection of COVID-19 based on chest X-rays using deep learning. In: Healthcare. 2022;10(2):343. https://doi.org/10.3390/healthcare10020343.
17. Erdaw Y, Tachbele E. Machine learning model applied on chest X-ray images enables automatic detection of COVID-19 cases with high accuracy. Int J Gen Med. 2021;14:4923–31. https://doi.org/10.2147/IJGM.S325609.
18. Ahmad HK, Milne MR, Buchlak QD, Ektas N, Sanderson G, Chamtie H, *et al.* Machine learning augmented interpretation of chest X-rays: A systematic review. Diagnostics. 2023;13(4):743. https://doi.org/10.3390/diagnostics13040743.
19. Kwekha-Rashid AS, Abduljabbar HN, Alhayani B. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. Appl Nanosci. 2023;13(3):2013–25. https://doi.org/10.1007/s13204-021-01868-7.
20. Heidari A, Jafari Navimipour N, Unal M, Toumaj S. Machine learning applications for COVID-19 outbreak management. Neural Comput Appl. 2022;34(18):15313–15348. https://doi.org/10.1007/s00521-022-07424-w.

21. Akhtar A, Akhtar S, Bakhtawar B, Kashif AA, Aziz N, Javeid MS. COVID-19 detection from CBC using machine learning techniques. Int J Technol Innov Manag. 2021;1(2):65–78. https://doi.org/10.54489/ijtim.v1i2.22.
22. Khoshbakhtian F, Ashraf AB, Khan SS. Covidomaly: A deep convolutional autoencoder approach for detecting early cases of covid-19. arXiv Prepr arXiv201002814. 2020.
23. Mahesh B. Machine learning algorithms. A Review. Int J Sci Res. 2020;9(1):381–6. http://dx.doi.org/10.21275/ART20203995.
24. Zaki SM, Jaber MM, Kashmoola MA. Diagnosing COVID-19 infection in chest X-Ray images using neural network. Baghdad Sci J. 2022;19(6):1356–61. https://doi.org/10.21123/bsj.2022.5965.
25. Eljamassi DF, Maghari AY. COVID-19 detection from chest X-ray scans using machine learning. Proc 2020. Int Conf Promis Electron Technol. ICPET 2020. 2020;1–4. https://doi.org/10.1109/ICPET51420.2020.00009.
26. Samsir S, Sitorus JHP, Ritonga Z, Nasution FA, Watrianthos R. Comparison of machine learning algorithms for chest X-ray image COVID-19 classification. J Phys Conf Ser. 2021;1933(1):012040. https://doi.org/10.1088/1742-6596/1933/1/012040.
27. Mohammad-Rahimi H, Nadimi M, Ghalyanchi-Langeroudi A, Taheri M, Ghafouri-Fard S. Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: A scoping review. Front Cardiovasc Med. 2021;8:638011. https://doi.org/10.3389/fcvm.2021.638011.
28. Zargari Khuzani A, Heidari M, Shariati SA. COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images. Sci Rep. 2021;11(1):9887. https://doi.org/10.1038/s41598-021-88807-2.
29. Johri S, Goyal M, Jain S, Baranwal M, Kumar V, Upadhyay R. A novel machine learning-based analytical framework for automatic detection of COVID-19 using chest X-ray images. Int J Imaging Syst Technol. 2021;31(3):1105–19. https://doi.org/10.1002/ima.22613.
30. Alafif T, Tehame AM, Bajaba S, Barnawi A, Zia S. Machine and deep learning towards COVID-19 diagnosis and treatment: Survey, challenges, and future directions. Int J Environ Res Public Health. 2021;18(3):1117. https://doi.org/10.3390/ijerph18031117.
31. Cavallo AU. Texture analysis in the evaluation of COVID-19 pneumonia in chest X-Ray images: A proof of concept study. Curr Med Imaging Rev. 2022;17(9):1094–102. https://doi.org/10.2174/1573405617999210112195450.
32. Ahmed Ali H, Hariri W, Smaoui Zghal N, Ben Aissa D. A Comparison of machine learning methods for best accuracy COVID-19 diagnosis using chest X-Ray Images. 2022 IEEE 9th Int Conf Sci Electron Technol Inf Telecommun SETIT 2022. 2022;(Ml):349–55. https://doi.org/10.1109/SETIT54465.2022.9875477.
33. Arif ZH, Cengiz K. Severity classification for COVID-19 infections based on lasso-logistic regression model. Int J Math Stat Comput Sci. 2022;1:25–32. https://doi.org/10.59543/ijmscs.v1i.7715.
34. Dara OA, Lopez-Guede JM, Raheem HI, Rahebi J, Zulueta E, Fernandez-Gamiz U. Alzheimer's disease diagnosis using machine learning: A survey. Appl Sci. 2023;13(14):8298. https://doi.org/10.3390/ijerph18031117.
35. Sheikh BU, Zafar A. Robust medical diagnosis: A novel two-phase deep learning framework for adversarial proof disease detection in radiology images. J Imaging Inform Med. 2024;37(1):308–338. https://doi.org/10.1007/s10278-023-00916-8.
36. Peng T, Wang Y, Xu TC, Chen X. Segmentation of lung in chest radiographs using hull and closed polygonal line method. IEEE Access. 2019;7:137794–810. https://doi.org/10.1109/access.2019.2941511.
37. Attallah O. A deep learning-based diagnostic tool for identifying various diseases via facial images. Digit Heal. 2022;8:20552076221124430. https://doi.org/10.1177/20552076221124432.
38. Zaman A, Khattak SS, Hassan Z. Medical imaging for the detection of tuberculosis using chest radio graphs. In: 2019 International Conference on Advances in the Emerging Computing Technologies (AECT). IEEE. 2020;1–5. https://doi.org/10.1109/aect47998.2020.9194212.
39. Issarti I, Consejo A, Jiménez-García M, Hershko S, Koppen C, Rozema JJ. Computer aided diagnosis for suspect keratoconus detection. Comput Biol Med. 2019;109:33–42. https://doi.org/10.1016/j.compbiomed.2019.04.024.
40. Alavijeh FS, Mahdavi-Nasab H. Multi-scale morphological image enhancement of chest radiographs by a hybrid scheme. J Med Signals Sens. 2015;5(1):5 9. PMID: 25709942.
41. Rim B, Kim J, Hong M. Gender classification from fingerprint-images using deep learning approach. In: Proceedings of the International Conference on Research in Adaptive and Convergent Systems. 2020;7–12. https://doi.org/10.1145/3400286.3418237.
42. Phung VH, Rhee EJ. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. Appl Sci. 2019;9(21):4500. https://doi.org/10.3390/app9214500.
43. Mhawi DN, Hashem SH. Proposed hybrid correlation feature selection forest panalized attribute approach to advance IDSs. Mod Sci. 2021;7:15. https://doi.org/10.33640/2405-609x.3166.
44. Ebied HM. Feature extraction using PCA and Kernel-PCA for face recognition. In: 2012 8th International Conference on Informatics and Systems (INFOS). IEEE. 2012. MM–72.
45. Karamizadeh S, Abdullah SM, Manaf AA, Zamani M, Hooman A. An overview of principal component analysis. J Signal Inf Process. 2013;4(3B):173. https://doi.org/10.4236/jsip.2013.43B031.
46. Poon B, Amin MA, Yan H. PCA based human face recognition with improved method for distorted images due to facial makeup. In: Proceedings of the international multi conference of engineers and computer scientists, Hong Kong. 2017.
47. Reza MS, Ma J. ICA and PCA integrated feature extraction for classification. In: 2016 IEEE 13th International Conference on Signal Processing (ICSP). IEEE. 2016;1083–8. https://doi.org/10.1109/icsp.2016.7877996.
48. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends. 2021;2(01):20–8. https://doi.org/10.38094/jastt20165.
49. Navada A, Ansari AN, Patil S, Sonkamble BA. Overview of use of decision tree algorithms in machine learning. Pro IEEE Control Syst Grad Res Colloquium, ICSGRC. 2011;37–42. https://doi.org/10.1109/ICSGRC.2011.5991826.
50. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
51. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22–27, 2010 Keynote, Invited and Contributed Papers. Springer. 2010;177–86. https://doi.org/10.1007/978-3-7908-2604-316.

52. Emon MU, Islam R, Keya MS, Zannat R. Performance analysis of chronic kidney disease through machine learning approaches. In: 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE. 2021;713–9. https://doi.org/10.1109/ICICT50816.2021.9358491.
53. Maalouf M. Logistic regression in data analysis: An overview. Int J Data Anal Tech Strateg. 2011;3(3):281–99. https://doi.org/10.1504/IJDATS.2011.041335.
54. Gupta H V, Kling H, Yilmaz KK, Martinez GF. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol. 2009;377(1–2):80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
55. Romadhon MR, Kurniawan F. A comparison of Naive Bayes methods, logistic regression and KNN for predicting healing of Covid-19 patients in Indonesia. In: 3rd east Indonesia conference on computer and information technology (eiconcit). IEEE. 2021;41–4. https://doi.org/10.1109/EIConCIT50028.2021.9431845.
56. Zhang M-L, Zhou Z-H. A k-nearest neighbor based algorithm for multi-label classification. In: IEEE International Conference on Granular Computing. IEEE. 2005;718–21. https://doi.org/10.1109/GRC.2005.1547385.
57. Ontivero-Ortega M, Lage-Castellanos A, Valente G, Goebel R, Valdes-Sosa M. Fast Gaussian Naïve Bayes for searchlight classification analysis. Neuroimage. 2017;163:471–9. https://doi.org/10.1016/j.neuroimage.2017.09.001.
58. Jahromi AH, Taheri M. A non-parametric mixture of Gaussian Naive Bayes classifiers based on local independent features. In: Artificial intelligence and signal processing conference (AISP). IEEE. 2017;209–12. https://doi.org/10.1109/AISP.2017.8324083.
59. Haghighi S, Jasemi M, Hessabi S, Zolanvari A. PyCM: Multiclass confusion matrix library in Python. J Open Source Softw. 2018;3(25):729. https://doi.org/10.21105/joss.00729.
60. Ali HA, Zghal NS, Hariri W, Aissa D Ben. Fast hybrid deep neural network for diagnosis of COVID-19 using chest X-Ray images. Int J Adv Comput Sci Appl. 2023;14(3):553–64. https://doi.org/10.14569/ijacsa.2023.0140364.
61. Haouli I-E, Hariri W, Seridi-Bouchelaghem H. COVID-Attention: Efficient COVID19 detection using pre-trained deep models based on vision transformers and X-ray images. Int J Artif Intell Tools. 2023;32(08):2350046. https://doi.org/10.1142/S021821302350046X.

مقارنة خوارزميات التعلم الآلي القائمة على تحليل المكونات الرئيسية لتصنيف كورونا باستخدام صور الصدر بالأشعة السينية

حسين احمد علي[1،2]، وليد حريري[3]، نادية سماوي[4]، دلندى بن عيسى[1]

[1]مختبر *ابحاث الموجات الميكرو الكترونية، الكلية العلوم، الجامعة تونس المنار، تونس المنار، تونس.*
[2]*كلية علوم الحاسبات، جامعة كركوك، كركوك، العراق.*
[3]مختبر *LABGED لعلوم الحاسوب، قسم الاعلام اللآلي، جامعة باجي مختار عنابة، الجزائر.*
[4]مختبر *التحكّم وإدارة الطاقة، المدرسة الوطنية للمهندسين بصفاقس، تونس.*

الخلاصة

أدى الانتشار السريع لجائحة كوفيد-19 إلى إجهاد أنظمة الرعاية الصحية العالمية، مما استلزم أساليب تشخيص فعالة. في حين أن تفاعل البوليميراز المتسلسل (PCR) واختبارات المستضدات شائعة، إلا أن لها حدودًا في السرعة والدقة. يعد تعزيز دقة تقنيات التصوير، وخاصة (الأشعة السينية للصدر) و(التصوير المقطعي المحوسب)، أمرًا بالغ الأهمية للكشف عن تشوهات الرئة المرتبطة بكوفيد-19. يُفضل استخدام الأشعة السينية للصدر، لكونه فعال من حيث التكلفة ويمكن الوصول إليه، على الأشعة المقطعية، لكن التشخيص الدقيق غالبًا ما يتطلب دعمًا تكنولوجيًا. ولمعالجة هذه المشكلة، تتوفر مجموعة بيانات شاملة لصور الأشعة السينية للصدر مصنفة إلى خمس فئات على Kaggle . تتضمن معالجة مثل هذه البيانات خطوات مثل تحويل (التدرج الرمادي، وضبط كثافة الصورة، وتغيير الحجم، واستخراج الميزات باستخدام تحليل المكونات الرئيسية). تقنيات التعلم الآلي، بما في ذلك (شجرة القرار)، و(الغابات العشوائية)، و(الانحدار التدرجي العشوائي)، و(الانحدار اللوجستي)، و(نظرية البايزي الساذج)، و (خوارزمية الجيران الاقرب)، هي تقنيات المستخدمة لتصنيف الصور. يُظهر (شجرة القرار) أعلى دقة بنسبة 88%، متفوقًا على النماذج الأخرى مثل (نظرية البايزي الساذج 77%)، و (خوارزمية الجيران الاقرب 71%)، (الانحدار التدرجي العشوائي 70%)، (الانحدار اللوجستي 74%)، الغابات العشوائية 45% شجرة القرار إنه يتفوق باستمرار عبر مقاييس التقييم مثل درجة الضبط، والحساسية، الدقة، والاستدعاء، بمتوسط مرجح بنسبة 88%. ومع ذلك، يعتمد اختيار النموذج الأمثل لتعلم الآلة على عوامل مثل خصائص مجموعة البيانات وتفاصيل التنفيذ. وبالتالي، يعد النظر بعناية في هذه العوامل أمرًا بالغ الأهمية عند اختيار نموذج تعلم الآلة لتشخيص كوفيد-19 عبر تصنيف صور الأشعة السينية للصدر.

الكلمات المفتاحية: صور الأشعة السينية للصدر، كوفيد-19، شجرة القرار، الانحدار التدرجي العشوائي، نظرية البايزي الساذج، التعلم الالي.