Deep Learning-based diagnosis of COVID-19 using Chest X-Ray Images

Salah Zaher
Business Administration Information System Department
Taibah University, Saudi Arabia
Cairo University, Egypt

ABSTRACT
COVID-19 is a highly contagious disease caused by SARS-CoV-2 and has tragically claimed over 6 million lives worldwide. Early detection is crucial for containing its spread and providing appropriate treatment. In this article, we present a system to help physicians diagnose COVID-19 pneumonia by analyzing chest radiographs (CXR). The proposed system uses a dataset of 7232 CXR images, where 80% were used for training and 20% for testing. The methodology used in the proposed system consists of four steps: preprocessing the images through image processing, extracting features from the images using a pre-trained VGG16 model as a deep learning technique for feature extraction, applying two machine learning algorithms, namely Random Forests and Support Vector Machines (SVMs), and finally comparing the performance of the two algorithms. The SVM classifier achieved a classification accuracy of 97.2% at Gamma = 0.001 and C = 5, whereas Random Forests achieved an accuracy of 87%. Thus, the SVM classifier proved to be more capable in this type of classification.

General Terms
Deep Learning, **Machine learning Algorithms**.

Keywords
Covid-19, SARS-CoV-2, X-ray (CXR), VGG16,deep learning, machine learning, support vector machines (SVMs), random forest.

1. INTRODUCTION
COVID-19 has had a catastrophic impact on the world where the case is more with more than 6.2 million deaths worldwide [1]. COVID-19 is a highly contagious infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2],which can spread very rapidly. Early detection of the disease is a very important factor to stop the spread of coronavirus because this virus is infectious and cannot be treated appropriately [3]. So far, polymerase chain reaction (RT-PCR) is a gold standard screening method for detecting COVID-19 cases is reverse transcriptase– which can detect the nucleic acid of the virus

from nasopharyngeal or oropharyngeal swabs[4]. However, RT-PCR can only detect 60–70% of true COVID-19 symptomatic patients [5][6][7].In addition, detecting an infected person is a very difficult task because the disease symptoms are very similar to many other diseases, the simplest of which is the common cold, pneumonia. Deep Learning (DL) techniques have become increasingly popular in diagnosing various diseases, especially in medical imaging[8][9][10]. These techniques automatically extract image features without the need for any intervention. This makes them particularly useful for classifying infections in chest X-ray images, including COVID-19.DL techniques are used to automatically analyze chest X-ray images to detect COVID-19 imaging in patient images. This real-time support is less costly, more time-efficient, and more accurate, helping physicians and health professionals provide better care to physicians in diagnosing COVID-19 pneumonia using CXR images.

2. RELATED WORK
The authors of [11] proposed a CNN model known as a deep learning model for X-ray-based COVID-19 classification (COV-SNET). The researchers' extended dataset included 3913 COVID-19 images, 7966 normal photos, and 5441 pneumonia images. Their unique approach is to use a huge number of X-ray pictures, yet their work only demonstrated 95% sensitivity and 75% to 80% accuracy for both multi-class and binary classification. The results alone are insufficient to provide an accurate diagnosis. In [12], The authors proposed a new CNN model for automatically diagnosing COVID-19 disease based on raw X-ray images. Their model predicted COVID-19 instances with 87.02% accuracy. In [13], the authors' TL model-based VGG16-Network achieved a multi-classification accuracy of 91.69% for cases infected with (i) coronavirus, (ii) normal, and (iii) pneumonia. The model in this research is based on X-ray pictures from an open-source dataset.

3. MACHINE LEARNING (ML)
The goal of machine learning is to extract knowledge from data. It is a research area that combines statistics, artificial intelligence, and computer science. It is also known as predictive analytics or statistical learning. Machinelearning approaches have become increasingly common in everyday life in recent years. [14].

3.1 Supervised Learning
The supervised learning[15] approach uses labeled data where the input and output are known. The algorithm is trained on the data set and modified until it reaches an acceptable level of performance. Examples of supervised learning problems include:

- Regression problems
- Classification problems

3.2 Unsupervised Learning
Unsupervised learning[16] is an approach where the output is unknown. In this method, algorithms learn on their own and discover an impressive structure in the data. The primary goal of these algorithms is to decipher the underlying distribution in the data to gain more knowledge about it. Unsupervised learning problems include clustering and association.

3.2.1 *Random Forest*
Random Forest is a popular supervised machine learning algorithm built from decision tree algorithms. A Random

Forest removes the limitations of the decision tree algorithm by reducing overfitting of data sets and increasing precision.
This algorithm can be used for both classification and regression problems in ML. It is based on the concept of group learning, where multiple classifiers are combined to solve a complex problem and improve model performance.

As the name implies, the Random Forest algorithm is a classifier that contains multiple decision trees on different subsets of a given data set and takes the average to improve the predictive accuracy of that data set. Instead of relying on a single decision tree, a Random Forest takes the prediction from each tree and relies on a majority vote for the predictions and predicts the final output. A larger number of trees in a forest leads to higher accuracy and avoids overfitting. [17]. Figure 1 is a schematic diagram that illustrates a random forest.

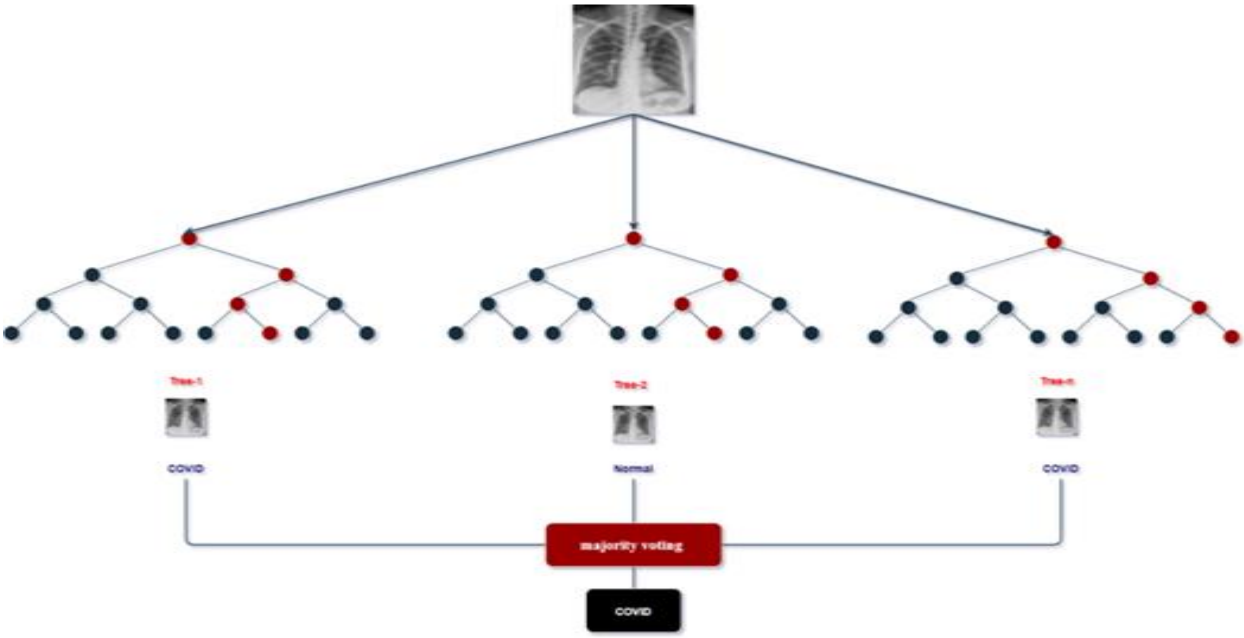

Fig.1. Random Forest Diagram

3.2.2 *Support Vector Machine (SVM)*
SVM[18] is an effective supervised machine learning algorithm that has gained popularity since its inception in 1992 by B.E. Boser et al. Its success in handwritten digit recognition in 1994 cemented its reputation as a powerful tool for both regression and classification problems [5], though it is especially renowned for its classification capabilities. SVM is widely regarded as one of the most exceptional off-the-shelf supervised .
Some basic terms used in the algorithm :The support vector , Hyper-Plane and Margin.

- Hyperplane:is a decision plane that separates between a set of objects with different class memberships.
- Support Vectors: are the data points closest to the hyperplane. These points better define the dividing line by calculating the edges. These points
- The 'C' parameter plays a crucial role in determining the trade-off between maximizing the margin and minimizing the misclassification of data points. When 'C' has a lower value, the algorithm creates a small margin, which in turn leads to a lower error rate. Conversely, when 'C' is set to a higher value, a larger margin is created, which can sometimes lead to misclassification of data points.
- The Gamma parameter
  The gamma value plays a crucial role in determining the performance of a model. A small gamma value implies that the model considers only a few data points that are close to the plausible margin line, leading to underfitting. Conversely, a large gamma value causes the model to involve a large number of data points, even those that are distant, in calculating the separation line. This can result in overfitting, which can negatively impact the model's performance [9].

4. DEEP LEARNING

Deep learning [19], which is derived from machine learning methods based on artificial neural networks with representation learning, is a form of artificial intelligence. It can learn from data using supervised, unsupervised, or semi-supervised learning techniques, and create patterns for use in decision-making. The ANN algorithm is considered the heart of deep learning due to its architecture, which is inspired by the actions of the human brain in processing data and creating patterns for use in decision-making. While the NN algorithm contains three layers, including the input layer, hidden layer, and output layer, deep learning algorithms consist of more than three layers, including the input layer, more than one hidden layer, and output layer. Figure 3 depicts the relationship between Deep Learning (DL) and Artificial Neural Networks (ANN). ANN is a branch of machine learning that focuses on developing and refining neural networks for the purpose of recognizing patterns in data. On the other hand, DL is a more advanced and sophisticated version of ANN that employs deep neural networks with multiple layers to process and analyze massive amounts of data. It is worth noting that since DL is a more advanced form of ANN, ANN is considered to be a subset of machine learning that incorporates DL techniques.

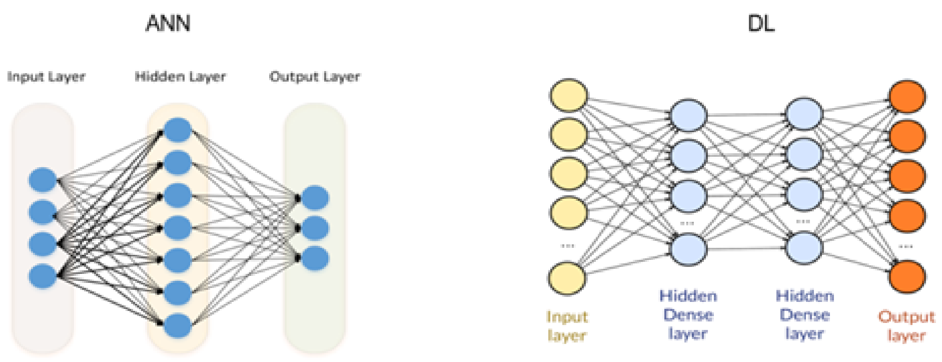

Figure 3: ANN and DL Relationship

4.1 Convolutional Neural Networks (CNN)

CNN[20], also known as ConvNet, is a type of feed-forward neural network that is widely used in tasks like image analysis, natural language processing, and other complex image classification problems. With its ability to pick out and detect patterns from images and text and make sense of them, CNN has become an essential tool for machine learning and artificial intelligence applications. By leveraging its advanced algorithms and deep learning capabilities, CNN can help organizations and researchers gain valuable insights into complex data sets and solve some of the most challenging problems in the field of computer science. The architecture of the convolutional neural network (CNN) is illustrated in Figure 4. Convolutional Neural Networks (CNNs) consist of several layers that are designed to extract features from input data such as images. The main layers in CNNs are:

- Input Layer: This is the first layer in the network that takes in the input data such as images.
- Convolutional Layer :This layer applies a set of learnable filters to the input data in a sliding window fashion, producing output feature maps. The filters are designed to detect various patterns and features in the input data such as edges, lines, and shapes.
- ReLU Layer: This layer applies the Rectified Linear Unit (ReLU) activation function to the output of the convolutional layer. ReLU helps to introduce non-linearity into the network and improve its ability to learn complex patterns.
- Pooling Layer :This layer reduces the spatial size of the output feature maps from the convolutional layer by down sampling the data. This helps to reduce the computation required in the network and make it more efficient.
- Fully Connected Layer :This layer connects all the neurons from the previous layer to the current layer, allowing the network to learn complex features and patterns from the input data.
- Output Layer :This is the final layer in the network that produces the output predictions based on the input data. The type of output layer depends on the task the network is designed to perform. For example, for classification tasks, the output layer is typically a softmax activation function that produces probabilities for each class. These layers work together to extract relevant features from the input data and produce accurate predictions for various computer vision tasks such as image classification, object detection, and segmentation

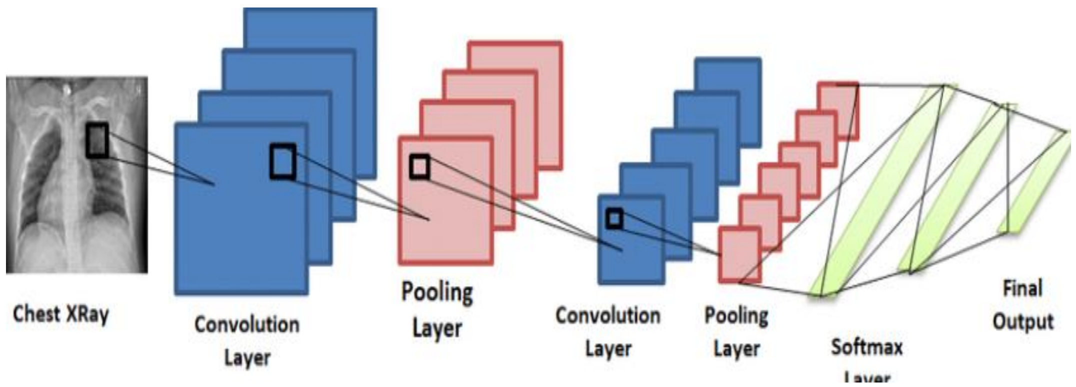

Figure 4: CNN Architecture

4.2 VGG16 Model

The VGG 16 [21]model is a convolutional neural network architecture that was introduced in 2014 by the Visual Geometry Group (VGG) at the University of Oxford. It contains 16 layers, including 13 convolutional layers and 3 fully connected layers. The VGG 16 model is widely used in computer vision tasks, such as image classification, object detection, and segmentation. It is known for its simplicity and high accuracy on various image recognition benchmarks.

4.2.1 *VGG16 model configuration*

As shown in Fig 5,VGG16 uses a 3x3 filter in convolution layers with a stride 1, and 2x2 filter padding and max pool layer with stride 2. It follows this convolution and max pool layers consistently throughout the entire architecture. Finally, it has 2 fully connected layers followed by a SoftMax for output[22] .

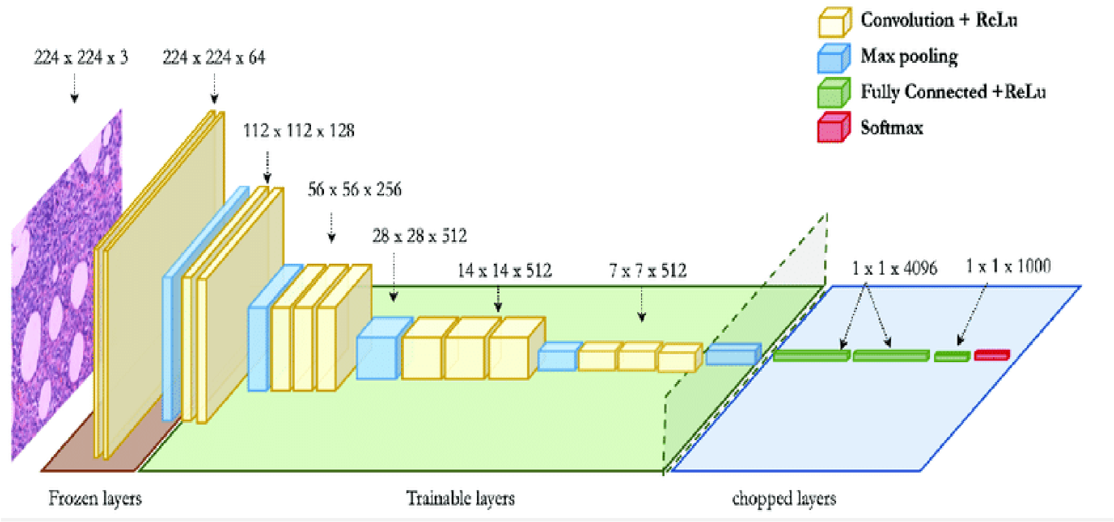

Fig5: VGG16 Configuration

4.2.2 Performance metrics

There are several metrics used to evaluate the performance of Convolutional Neural Networks (CNNs). Some of the most commonly used metrics include:

- Accuracy: This metric measures the percentage of correctly classified images in the dataset. It is calculated by dividing the number of correctly classified images by the total number of images in the dataset.

$$Accuracy = \frac{TP + FN}{TP + FN + FP + TN} \quad (1)$$

- Precision: This metric measures the percentage of correctly classified positive instances out of all instances classified as positive. It is calculated by dividing the number of true positives by the sum of true positives and false positives.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- Recall: This metric measures the percentage of correctly classified positive instances out of all actual positive instances in the dataset. It is calculated by dividing the number of true positives by the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

- F1 score: This metric is a weighted average of precision and recall, and is often used when both precision and recall are important.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

- Confusion matrix:
  Fig 6 shows the confusion matrix, which displays the number of true positives, false positives, true negatives, and false negatives for each class in the dataset. This visualization technique is commonly used to evaluate the performance of a CNN and to identify areas where it may be making errors. Where:
  TP: True Positives
  FP: False Positives
  TN: True Negatives
  FN: False Negatives
  These metrics can be used to evaluate the performance of a CNN on a variety of tasks,

5. PROPOSED SYSTEM DESIGN AND METHODOLOGY

The system's architectural design provides an overview of the different components and how they interact with each other. Figure 4.1 displays the architectural design of the diagnosis system for COVID-19 using Chest X-Ray Images. As the proposed system follows a deep learning approach, the first phase involves training the model to classify CXR images in the given dataset. In phase two, the user interacts with the system by inputting the CXR image for diagnosis and receives the detected disease back from the system. The research methodology that was referenced contains a total of eight phases that are designed to guide the development of a system. Each phase must be completed before moving on to the next one, and there is no overlapping between the phases. The output of one phase serves as the input for the next phase sequentially. The proposed system's process diagrams are depicted in Figure 8.

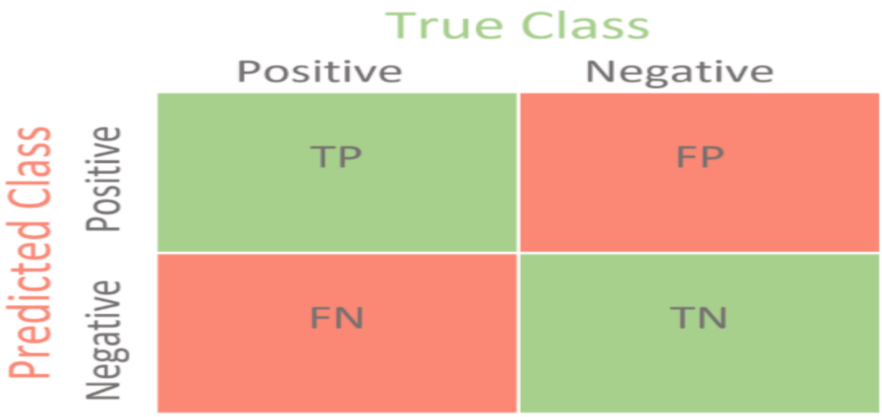

Fig 6: Confusion Matrix

5.1 Data set

The proposed system will be utilizing the COVID-19 Radiography Database, which consists of chest X-ray images. The dataset was downloaded from Kaggle [23] and contains 3616 COVID-19 positive cases along with 3616 Normal cases. Each image is 256 X 256 pixels in size. Figure 9 displays an example of COVID-19 CXR images from the dataset.

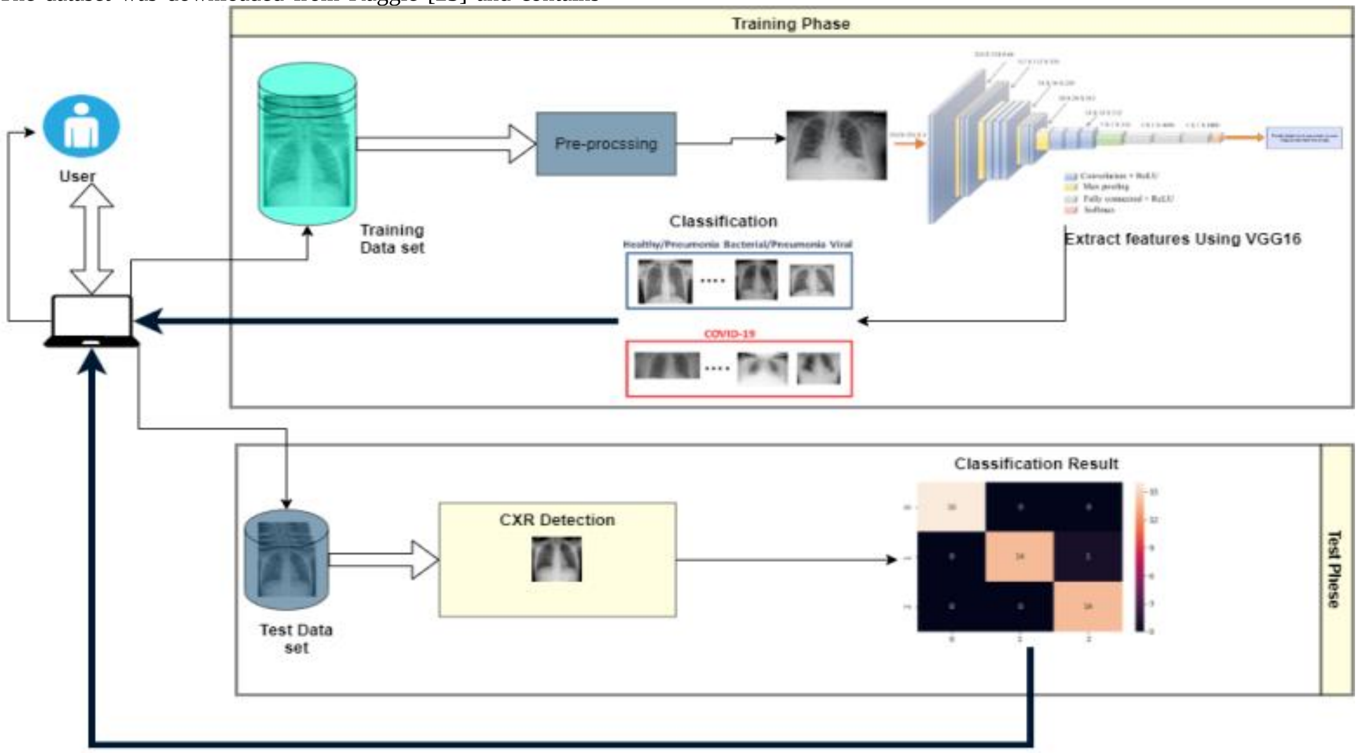

Fig 7: The system architectural diagram

Fig 8: The proposed system methodology with eight phases

5.2 Preprocessing
The dataset images undergo pre-processing in this phase, which involves resizing the images from 256X256 to 224 X

5.3 Pre-training model
In this phase, the dataset was trained using a VGG16 pre-training model in order to extract CXR image features.

5.4 VGG16 tuning
The tuning of the VGG16 model to meet the requirements of the proposed system has been finished in this phase.

5.5 Extract CXR image features
After fine-tuning the pre-training model, the features of the x-ray images are extracted in this phase.

5.6 RF Versus SVM Classifiers
In machine learning, RF (Random Forest) and SVM (Support Vector Machine) classifiers are two commonly used algorithms for classification tasks. RF is an ensemble learning method that consists of multiple decision trees, whereas SVM is a binary linear classifier that constructs a hyperplane to separate classes. While both algorithms have their advantages and disadvantages, RF is generally preferred for complex datasets with high dimensionality, while SVM is more suitable for datasets with smaller feature sets and a clear separation between classes. Ultimately, the choice between RF and SVM depends on the specific problem at hand and the characteristics of the dataset.The proposed system incorporates the use of random forest and SVM classifiers to generate results. The generated results are then compared, and the one with better accuracy is selected.

5.7 Training model
During this phase, the classifier was trained on the proposed dataset which was divided into 80% for training and 20% for validation. SVM parameters were chosen as Gamma=0.001 C=5 for random forest n-estimators=200, Random-state=30.

5.8 Model Evaluation
In this phase, the classifier will be evaluated using various performance metrics to determine its effectiveness and accuracy.

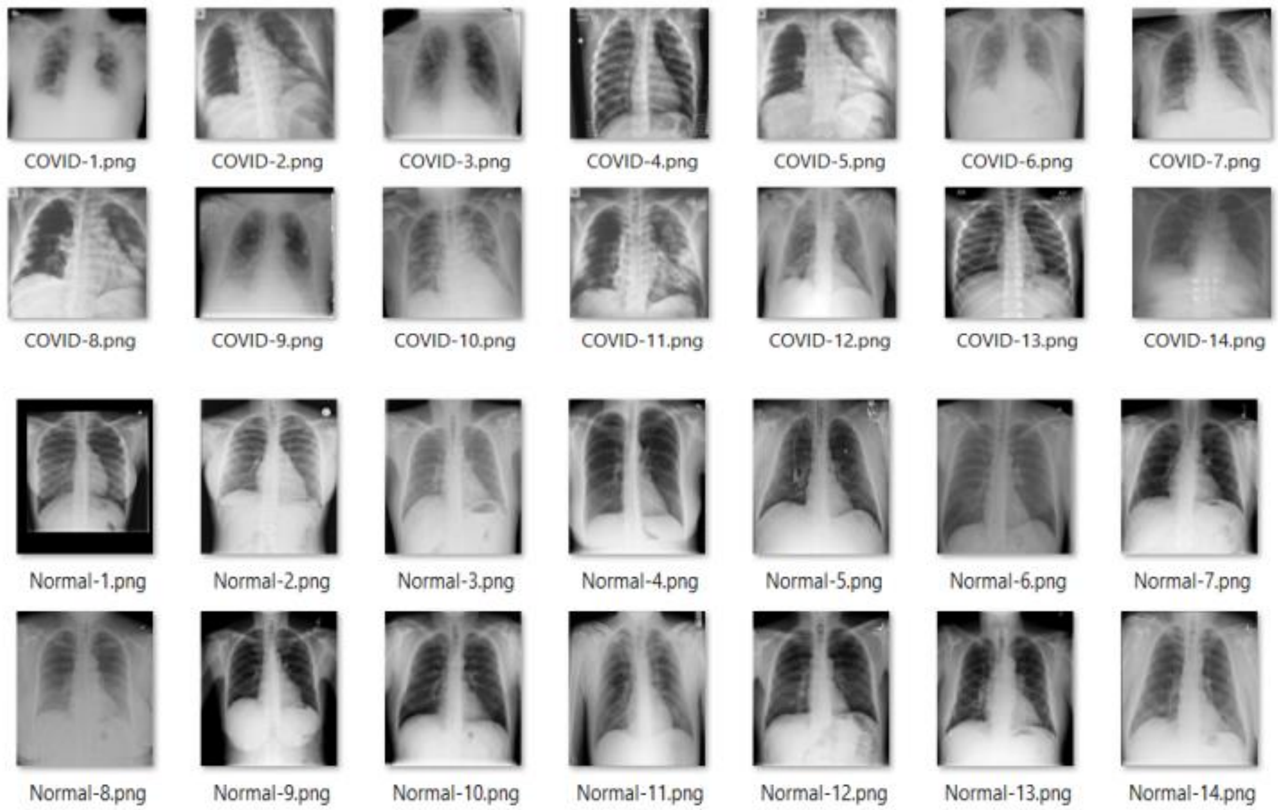

Fig 9: examples of COVID-19 and normal CXR images

6. TOOL USED:
- Google Colab environment

It is Python development environment, that run using Google cloud and provides free GPU, because training models on CPU takes a long time so using GPU is better than CPU. It allows us to write and execute our deep learning applications. Dealing with large datasets in Google Colab is easy because it offer uploading the dataset on Google drive and then map it to Google Colab using "Mount Drive" button. In addition, it is suitable for machine learning. It provides a lot of popular libraries such as TensorFlow, Keras, PyTorch, and OpenCV libraries.

- TensorFlow library

TensorFlow is open-source library used to implement deep learning and machine learning
models. It is used for fasr numerical computation also it uses computational graphs for data flow. The name TensorFlow means data flows through the graph, and each graph consists of nodes and each node represents mathematical operations. It's job is to transfer complex data to artificial neural network later for analysis. Moreover, TensorFlow advantage is that it can be deploy in a verity of platforms such as Android. Some of important features are true portability that means it runs on CPU, GPU, mobile computing . Also, language options it

written in C++ and Python as we use in this project. TensorBoard is used as visualization tool with TensorFlow library. Meny data types can be visualized using TensorBoard such as graphs, scales, images, histograms, and audio. Usually TensorBoard installed along with TensorFlow .

- Keras library
  Keras is another open-source library used for deep learning. It is high level neural network Application Programming Interface (API), it can be used alone or also used with TensorFlow library.One of it's benefits is that it can run easly on both CPU and GPU. In addition, it allow us to easily build neural networks. About TensorFlow, Keras is usually used in top of Tensorflow library.

7. RESULTS

In this study six experiments were performed on the random forest technique in this study, where we initially examined three alternative values for n_estimators, with the finding that 200_estimators achieved the best accuracy. In the following stage, we added a random_state to the best result, and the algorithm obtained the highest degree of accuracy by 86.7% at n_estimators =200 with random_state = 30. Table 1 displays the results of the random forest trials. Concerning the SVM algorithm experiment, we first compare the linear kernel with the rbf kernel. The linear kernel achieved the maximum accuracy of 95.4% in this trial. So we tried SVM with a linear kernel with varied C values, but all experiments produced the same degree of accuracy, as shown in Table 2. To do this, we modified the kernel to rbf with the addition of gamma, where we acquired the maximum accuracy from the parameter gamma, and then we added it to it with a C parameter to obtain the highest accuracy by 97.2% at kernel = rdf and C = 0.001, as shown in Table 3. Figure 10 shows the RBF kernel settings versus the accuracy.As shown in Table 4, when comparing SVM and RF, SVM achieved the highest accuracy (97%), while random forest achieved 87%. So our model used SVM as a classifier.

8. CONCLUSION

According to the findings of this study, artificial intelligence plays a critical role in achieving proactive and optimum COVID-19 quarantine responses. We proposed a model for categorizing COVID-19 CXR images using domain extension transfer learning in this research. The proposed system was effective in categorizing COVID-19 CXR images after pre-training and fine-tuning methods in the target domain. Using the SVM classifier with kernel = rbf and C = 0.001, the suggested technique captured COVID-19 with 97.2% accuracy.

In the future,As medical technology continues to advance, new methods of classifying diseases are emerging. COVID-19 is just one example of a respiratory illness that can benefit from these advancements. By utilizing the latest machine learning techniques, algorithms can be developed to accurately identify various types of lung diseases based on symptoms and medical history. This can have a significant impact on patient outcomes, allowing doctors and healthcare professionals to provide more targeted diagnoses and treatments.

- Competing Interest and Funding

I hereby state that I do not have any competing interests. Furthermore, I confirm that I am not receiving any funds from any individual or organization.

Table1: Random Forest Trial Results

Trial	kernel	Gamma	C	curacy
Trial 1	Linear	-	0.1	95.4%
Trial 2	Linear	-	1	95.4%
Trial 3	Linear	-	5	95.4%
Trial 4	Linear	-	10	95.4%

Table2: SVM with linear kernel Trials Results

Trials	n_estimators	random_state	Accuracy
Trial 1	150	10	85.0%
Trial 2	200	10	86.4%
Trial 3	250	10	86.5%
Trial 4	200	20	86.7%
Trial 5	200	30	86.8%
Trial 6	200	50	86.5%

Table3: SVM with rbf kernel Trials Results

Trial	kernel	Gamma	C	Accuracy
Trial 1	Rbf	0.001	1	96.1%
Trial 2	Rbf	0.01	1	84.1%
Trial 3	Rbf	1	1	84.1%
Trial 4	Rbf	0.001	0.1	87.0%
Trial 5	Rbf	0.001	5	97.2%
Trial 6	Rbf	0.001	10	87.0%

Table 4: SVM Vs. Random Forest Results

Algorithm/Metric	precision	recall	F1-score	Accuracy
SVM	97%	97%	97%	97.2%
Random Forest	87%	87%	87%	87%

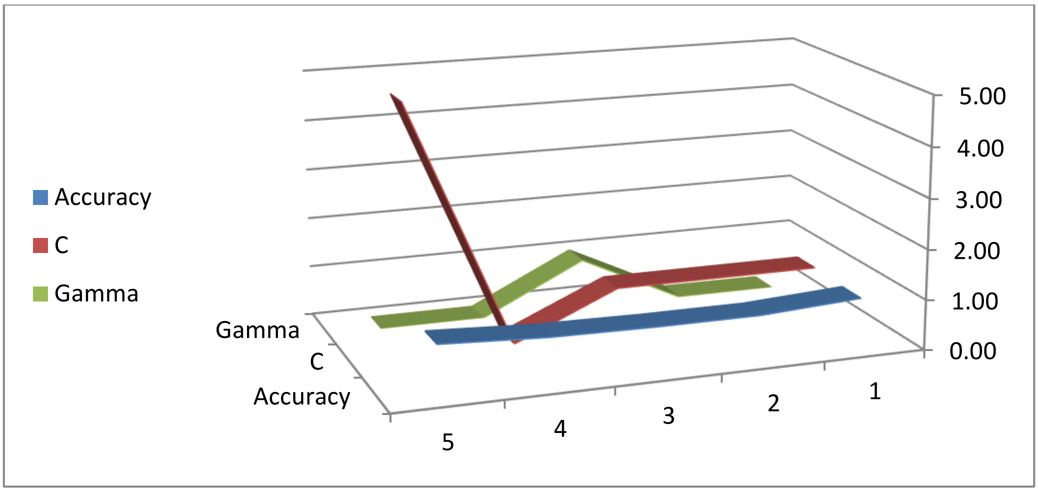

Fig 10 RBF kernel parameters versus accuracy

9. REFERENCES

[1] World Health Organization,28May2022. [Online]. Available: https://covid19.who.int/Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Sharma A, Ahmad Farouk I, Lal SK. COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission,Detection,Control and prevention viruses. 2021 Jan 29;13(2)

[3] Melika Lotfi,a,b Michael R. Hamblin,c,d,e and Nima Rezaeif COVID-19: Transmission, prevention, and potential therapeutic opportunities. Clin Chim Acta. 2020 Sep; 508: 254–266.

[4] Manas K. Akmatov, Nadine Koch, Marius Vital, Wolfgang Ahrens, Dieter Flesch-Janys, Julia Fricke, Anja Gatzemeier, Halina Greiser, Kathrin Günther, Michels, Susanne Moebus, Alexandra Nieters, Nadia 2017 May 12. doi: 10.1038/s41598-017-01212-6 PMCID:PMC5431815

[5] Ai Tao, Yang Zhenlu, Hou Hongyan, Zhan Chenao, Chen Chong, Lv Wenzhi, Tao Qian, Sun Ziyong, Xia Liming (2020) Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: A report of 1014 cases. Radiology 296(2):E32–E240. https ://doi.org/10.1148/radiol.2020200642

[6] Fang Yicheng, Zhang Huangqi, Xie Jicheng, Lin Minjie, Ying Lingjun, Pang Peipei, Ji Wenbin (2020) Sensitivity of chest ct for covid-19: Comparison to rt-pcr. Radiology 296(2):E115–E117 https://doi.org/10.1148/ radiol. 2020200432.

[7] Kanne Jeffrey P, Little Brent P, Chung Jonathan H, Elicker Brett M, Ketai Loren H (2020) Essentials for radiologists on covid-19: An update–radiology scientific expert panel. Radiology 296(2):E113–E114. https://doi.org/10.1148/radiol.2020200527

[8] C. Zheng et al., "Deep learning-based detection for COVID-19 from chest CT using weak label", 2020.

[9] Y. Cao et al., "Longitudinal assessment of COVID-19 using a deep learning-based quantitative CT pipeline: Illustration of two cases", Radiology Cardiothoracic Imag., vol. 2, 2020.

[10] [10]L. Huang et al., "Serial quantitative chest CT assessment of COVID-19: Deep-learning approach", Radiology Cardiothoracic Imag., vol. 2, 2020.

[11] . R. Hertel and R. Benlamri, "COV-SNET: A deep learning model for X-ray-based COVID-19 classification,"Informatics in Medicine Unlocked, vol. 24, no. 11, pp. 620-642, 2021.

[12] X. Qi, L. G. Brown, D. J. Foran, J. Nosher and I. Hacihaliloglu, "Chest X-ray image phase features for improved diagnosis of COVID-19 using convolutional neural network," International Journal for Computer Assisted Radiology and Surgery, vol. 16, no. 2, pp. 197–206, 2021.

[13] Hasan, M.D.K.; Ahmed, S.; Abdullah, Z.M.E.; Monirujjaman Khan, M.; Anand, D.; Singh, A.; AlZain, M.; Masud, M. Deep Learning Approaches for Detecting Pneumonia in COVID-19 Patients by Analyzing Chest X-Ray Images. Math. Probl. Eng. 2021,2021, 9929274.

[14] Andreas C. Müller, Sarah Guido Released September 2016 Publisher(s): O'Reilly Media, Inc.ISBN: 9781449369897"Introduction to Machine Learning with Python".

[15] Shakti Chourasiya, Suvrat Jain " A Study Review On Supervised Machine Learning Algorithms".International Journal of Computer Science and Engineering 2019by SSRG - IJCSE Journal Volume 6 Issue 8. doi: 10.14445/23488387/IJCSE-V6I8P104.

[16] Zoubin Ghahramani: Advanced Lectures on Machine Learning :ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures

[17] Brieman L, Random Forests, Machine Learning, 45,5-32, (2001) © 2001 Kluwer Academic Publishers. Manufactured in The Netherlands.

[18] Mariette Awad, Rahul Khanna:"Support Vector Machines for Classification"

[19] April 2015 DOI: 10.1007/978-1-4302-5990-9_3

[20] Shrestha, A.; Mahmood, A. Review of Deep Learning Algorithms and Architectures. IEEE Access 2019, 7, 53040–53065.

[21] Swapna, M.; Sharma, Y.K.; Prasad, B. CNN

Architectures: Alex Net, Le Net, VGG, Google Net, Res Net. Int. J. Recent Technol. Eng. 2020, 8, 953–959.

[22] Srikanth Tammina: Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images : International Journal of Scientific and Research Publications, Volume 9, Issue 10, October 2019 143 ISSN 2250-3153.

[23] Brian Kenji Iwana; Ryohei Kuroki; Seiichi chidaExplaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation :IEEE Xplore: Published in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) DOI: 10.1109/ICCVW.2019.00513

[24] COVID-19 Chest X-ray images and Lung masks Database available at: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/