Journal of Imaging Informatics in Medicine (2024) 37:308–338
https://doi.org/10.1007/s10278-023-00916-8

Robust Medical Diagnosis: A Novel Two-Phase Deep Learning Framework for Adversarial Proof Disease Detection in Radiology Images

Sheikh Burhan ul haque[1] · **Aasim Zafar**[1]

Received: 31 May 2023 / Revised: 23 September 2023 / Accepted: 8 October 2023 / Published online: 10 January 2024

Abstract

In the realm of medical diagnostics, the utilization of deep learning techniques, notably in the context of radiology images, has emerged as a transformative force. The significance of artificial intelligence (AI), specifically machine learning (ML) and deep learning (DL), lies in their capacity to rapidly and accurately diagnose diseases from radiology images. This capability has been particularly vital during the COVID-19 pandemic, where rapid and precise diagnosis played a pivotal role in managing the spread of the virus. DL models, trained on vast datasets of radiology images, have showcased remarkable proficiency in distinguishing between normal and COVID-19-affected cases, offering a ray of hope amidst the crisis. However, as with any technological advancement, vulnerabilities emerge. Deep learning-based diagnostic models, although proficient, are not immune to adversarial attacks. These attacks, characterized by carefully crafted perturbations to input data, can potentially disrupt the models' decision-making processes. In the medical context, such vulnerabilities could have dire consequences, leading to misdiagnoses and compromised patient care. To address this, we propose a two-phase defense framework that combines advanced adversarial learning and adversarial image filtering techniques. We use a modified adversarial learning algorithm to enhance the model's resilience against adversarial examples during the training phase. During the inference phase, we apply JPEG compression to mitigate perturbations that cause misclassification. We evaluate our approach on three models based on ResNet-50, VGG-16, and Inception-V3. These models perform exceptionally in classifying radiology images (X-ray and CT) of lung regions into normal, pneumonia, and COVID-19 pneumonia categories. We then assess the vulnerability of these models to three targeted adversarial attacks: fast gradient sign method (FGSM), projected gradient descent (PGD), and basic iterative method (BIM). The results show a significant drop in model performance after the attacks. However, our defense framework greatly improves the models' resistance to adversarial attacks, maintaining high accuracy on adversarial examples. Importantly, our framework ensures the reliability of the models in diagnosing COVID-19 from clean images.

Keywords Medical images · Adversarial examples · Robustness · Deep learning · COVID-19

Introduction

During the pandemic, the World Health Organization (WHO) considered RT-PCR the gold standard for diagnosing the virus. However, in many cases, COVID-19 patients remained undiagnosed due to the low sensitivity of RT-PCR [1–4], i.e., the false-negative rate is high. Failure to diagnose the disease at an early stage results in the patient not receiving adequate treatment on time, and due to the virus's highly infectious nature, the danger of the sickness spreading to a broader population increases. Sometimes, RT-PCR tests should be done numerous times for specific individuals to diagnose COVID-19 [5, 6]. The testing technique is costly and necessitates a complex manual process. The test findings take a long time to get, and there is a significant risk of healthcare staff becoming infected with the disease during the test. Moreover, sufficient training is essential for health professionals collecting samples for PCR. All of these limitations suggest that other rapid, accurate, and reliable diagnostic methods should be performed in addition to the RT-PCR test.

✉ Sheikh Burhan ul haque
sbuhaque@myamu.ac.in; shiekhburhan2013@gmail.com

Aasim Zafar
azafar.cs@amu.ac.in

[1] Department of Computer Science, Aligarh Muslim University, Uttar Pradesh, Aligarh 202002, India

Computer-aided diagnosis based on artificial intelligence technologies, including DL ML, has been rapidly expanding and active for the past ten years. Researchers have successfully employed DL techniques in disease diagnoses, such as cancer detection [7–9], Alzheimer's detection[10–13], and heart disease [14]. AI has also proven useful during the pandemic [15–18]. DL is the most promising and extensively used ML technology for disease diagnosis from medical images, such as radiology images. Due to the success of deep learning in the medical field, researchers successfully adopted AI technologies such as DL techniques to enhance the diagnosis of COVID-19 by using radiology images such as X-rays and CT images of lung regions. Most researchers have based their diagnosis models on transfer learning [19–26], while some have developed novel architectures [27–33]. Researchers have employed DL models to forecast the advancement and severity of COVID-19 in infected individuals [34–41].

Additionally, some studies have focused on analyzing past medical records to assess the likelihood of contracting COVID-19 [42–44]. Testing using deep models is straightforward, cost-effective, time-efficient, and accurate. Moreover, CT and X-ray images are easily accessible compared to RT-PCR test kits. Undoubtedly, COVID-19 detection models based on DL have performed exceptionally well in this domain, have a huge potential in medical science, and may one day be a standard method to diagnose COVID-19.

However, the deep learning-based model comes with certain inherent flaws. They are susceptible to various security attacks during their training and inferencing phase, forcing the model to make misclassifications [45]. During training, the model can be attacked by backdoor attacks or data poisoning attacks [46], in which the attacker either mislabels the training data or crafts a trigger and patches it to the training data sample. The attacker then trains the model on the malicious data samples, and the final model is contaminated. While as during inferencing, the model can be attacked by adding intelligently crafted perturbation to the simple image that causes the model to misclassify the simple image. This attack is called an adversarial attack [47]. In Fig. 1, we have demonstrated the possible security threats in the context of deep COVID-19 models.

This article highlights the issue of adversarial attacks on DL models used in deep COVID-19 diagnosis models. While several DL-based COVID-19 diagnosis systems have been proposed, very few have addressed the potential vulnerabilities of the system or how to mitigate them. We developed three deep learning models for COVID-19 diagnosis using radiology images (CT and X-ray images) by leveraging transfer learning with ResNet50 [48], VGG16 [49], and InceptionV3 [50] architectures, which classify the radiology images into non-COVID-19 pneumonia, e.g., (bacterial and viral pneumonia), normal (no pneumonia) and COVID-19 viral pneumonia classes. Despite the high performance of these models, we discovered their susceptibility to adversarial attacks, specifically FGSM [51], PGD [52], and BIM [53]. To fortify these models, we proposed a two-phase defense strategy. The first phase involved a modified adversarial training approach during the model training stage, where we trained the model on a random subset of the data and included multiple adversarial images of the same original image, each subjected to different image transformations such as rotation, brightness adjustment, and contrast variation. This approach enhanced the model's resilience by exposing it to a broader spectrum of potential adversarial examples. In the inference phase, we

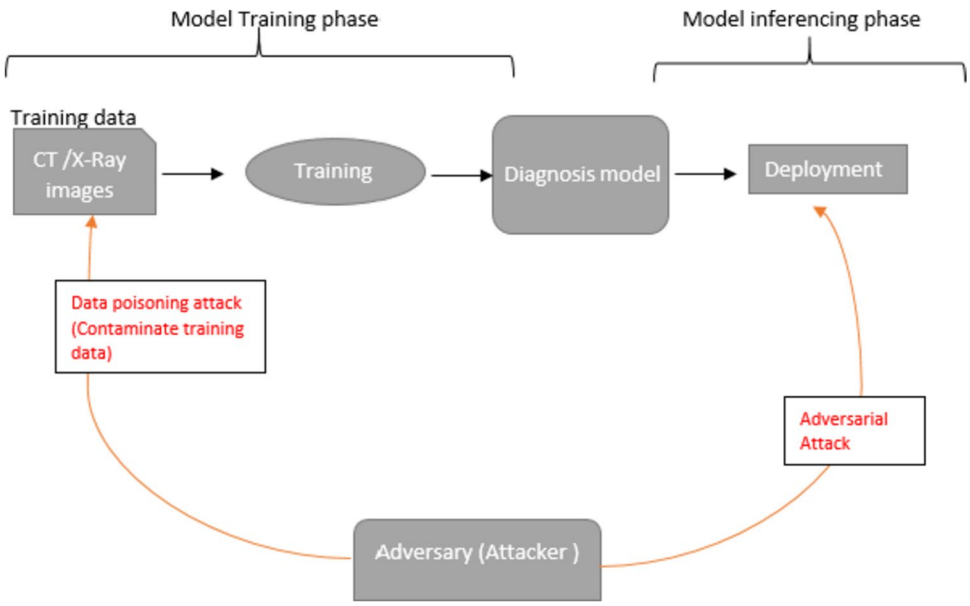

Fig. 1 Various attacks on deep COVID-19 diagnosis models, organized according to their respective phases

applied JPEG compression [54] to eliminate adversarial noise from the input data to reinforce the model's ability to withstand attacks. Through this two-phase defense strategy, our research aims to provide models that show promise in maintaining their performance and reliability throughout both training and real-world deployment stages, potentially serving as a valuable complement to biochemical tests.

Motivation

The motivations for conducting adversarial attacks on COVID-19 diagnosis systems are multifaceted. Adversarial attacks on COVID-19 diagnosis models are motivated by various factors. Malicious intent drives attackers to disrupt the accuracy of models, leading to misclassification of COVID-19 cases and sowing confusion among healthcare professionals and patients. Economic gain is another driver, as manipulating model outputs can benefit certain entities, such as pharmaceutical companies, by influencing perceptions of treatment efficacy. Social engineering and misinformation play a role, with adversaries exploiting model vulnerabilities to spread false information and undermine public trust in the healthcare system. Attacks also occur for research purposes, aiming to expose weaknesses in existing models or showcase alternative methodologies. Additionally, some attackers view manipulating COVID-19 diagnosis models as a technical challenge to test the robustness of AI systems. Addressing these motivations makes it possible to build more robust and resilient AI systems that can withstand adversarial attacks and maintain trust and accuracy in critical healthcare applications.

Contribution

The main contribution of the paper is listed below:

- Our study introduces and evaluates three deep COVID-19 diagnosis models using transfer learning techniques. We achieved remarkable accuracy rates on unseen clean data samples by leveraging the VGG-16, ResNet-50, and Inception-V3 architectures as base models. Specifically, on the X-Ray dataset, our models attained accuracy rates of 93.34%, 91.45%, and 94.65% for ResNet-50, VGG-16, and Inception-V3, respectively. Moreover, our models achieved 95.72%, 92.89%, and 95.03% accuracy on CT images for the corresponding architectures mentioned.
- We demonstrated that the deep COVID-19 diagnosis models based on conventional ML or DL approaches are susceptible to adversarial attacks by exposing them to the PGD, FGSM, and BIM attacks.
- A novel two-phase defense strategy was proposed to protect deep learning-based COVID-19 diagnosis models from adversarial attacks. The first phase (training) incorporates a sophisticated adversarial learning algorithm during the model training stage to enhance the model's robustness against adversarial examples. In the inference phase, JPEG compression was employed to remove adversarial noise from the input data, thereby improving the model's resilience against adversarial attacks.
- An extensive evaluation of the proposed defense mechanism was conducted, demonstrating its effectiveness in mitigating the adverse effects of adversarial attacks, thus ensuring the models' reliability in diagnosing COVID-19 from radiology images.

The paper is structured as follows. It begins with an introduction, providing an overview of the topic. "Motivation" presents a comprehensive literature review, discussing relevant prior research and studies. In "Contribution", the authors describe the dataset used in this study and explain the methodology employed. The methodology is divided into three parts: training COVID-19 diagnosis models, attacking the trained models, and implementing a secure two-phase defense approach. Moving on to "Related Work", the authors delve into the experimental results and outputs, providing a thorough analysis and interpretation of the findings. "COVID-19 Diagnosis Models" focuses on the conclusion, summarizing the main points discussed throughout the paper. Finally, the authors highlight the study's limitations and propose potential areas for future research and exploration.

Related Work

Initially, we comprehensively reviewed the current literature concerning COVID-19 diagnosis models that utilized deep learning and hybrid methodologies. Subsequently, we explored existing literature on adversarial attacks targeting deep COVID-19 models and strategies for defending against such attacks.

COVID-19 Diagnosis Models

Amidst the ongoing epidemic, there has been a surge of interest in projects aiming to develop effective COVID-19 diagnosis models. Many researchers have turned to convolutional neural networks (CNNs) for their remarkable performance in extracting valuable features from image data. Some studies have also explored hybrid approaches that combine ML methodologies with or without DL.

Recent advancements in COVID-19 diagnosis have been shaped by innovative ML and DL techniques. In one study [55], a Convolutional Neural Network (CNN) model demonstrated impressive capabilities by achieving a 98%

accuracy in distinguishing COVID-19 cases from healthy chest X-rays. Another significant development, COVID-AL [56], introduced a weakly-supervised deep active learning framework that efficiently diagnoses COVID-19 using CT scans, with over 95% accuracy even when trained on just 30% of labeled data.

Furthermore, research in [57] introduced the UC-MIL and BA-GCN models, revolutionizing COVID-19 diagnosis from chest CT scans by enhancing reliability and accuracy. The dual-branch combination network (DCN) [58] also excelled in COVID-19 diagnosis using chest CT scans, surpassing other models, particularly in detecting subtle lesions. Another study [59] focused on a computer-aided detection system for COVID-19 using chest X-rays, achieving high scores and interpretability.

Additionally, a robust multi-feature CNN [60] demonstrated remarkable accuracy in COVID-19 detection from chest X-rays, showcasing the potential of these techniques for fast and reliable diagnosis. Furthermore, an ensemble of CNNs [61] emphasized the power of combining models to enhance precision, recall, and accuracy when diagnosing COVID-19 from CT scans. Beyond imaging, [62] explored machine learning models based on routine blood exams as a viable alternative for screening COVID-19, achieving exceptional accuracy rates.

Moreover, [63] delved into artificial intelligence-based COVID-19 diagnosis from CT images, presenting promising results. In lung CT image-based diagnosis, an optimized deep convolutional neural network (DCNN) model named DCNN-IPSCA [64] excelled, achieving remarkable accuracy and speed. Finally, [65] employed CNN models to distinguish COVID-19 and other infections in lung X-ray scans, achieving high accuracy, and effectively used an LSTM model for forecasting COVID-19 cases in Italy with remarkable precision. Collectively, these diverse approaches represent significant progress in COVID-19 diagnostic capabilities, contributing to improved pandemic management and healthcare.

Adversarial Attack and Deep COVID-19 Systems

This section delves into the literature on adversarial attacks aimed explicitly at COVID-19 monitoring models. We examine the potential vulnerabilities of these models and their susceptibility to deliberate attacks through the manipulation of input data.

The realm of artificial intelligence has faced growing concerns over the vulnerability of neural networks to adversarial attacks, as underscored in [66]. This issue has prompted a surge in research endeavors dedicated to developing novel attack methods and mitigation strategies. Surprisingly, this line of inquiry has yet to be adequately explored concerning computer-based technologies designed to combat the COVID-19 pandemic, creating an urgent need for further scholarly investigation in this critical domain. The study in [67] delved into the susceptibility of the COVID-Net architecture to universal adversarial perturbation (UAP) attacks, revealing vulnerabilities in a system designed for COVID-19 patient classification using chest X-ray images. Meanwhile, [69] examined the susceptibility of DL algorithms utilized in medical Internet of Things (IoT) applications for COVID-19 diagnostics to adversarial attacks, shedding light on security vulnerabilities. This research explored various adversarial attack methods, including MI-FGSM, FGSM, L-BFGS, Deepfool, BIM, Carlini and Wagner (C&W), PGD, Foolbox, and Jacobian-based Saliency Map Attack (JSMA) [70].

Furthermore, the study outlined in [71] introduced a method to generate unrestricted adversarial examples targeting face recognition systems, showcasing its effectiveness in both white-box and black-box settings. This approach had a success rate of approximately 90% in deceiving target face recognition models, underscoring the urgency of developing robust defenses against adversarial attacks. In [72], a COVID-19 diagnosis model based on transfer learning faced increased vulnerability to adversarial attacks as perturbation levels escalated, impacting the accuracy of X-ray and CT imaging classification models. Another study [73] proposed an image-based medical adversarial attack method that consistently generated perturbations on medical images, with a focus on maximizing the deviation loss term while minimizing loss stabilization. Additionally, [74] explored passive and active attack methods on deep neural networks (DNNs) applied to medical datasets, revealing vulnerabilities in DNN inference engines used for COVID-19 detection from chest X-ray images. Lastly, [75] investigated the impact of adversarial attacks on widely employed neural network architectures, highlighting the varying vulnerabilities of different models to these attacks and emphasizing the need for robust defense mechanisms. These collective research efforts underscore the significance of addressing adversarial vulnerabilities in AI systems, particularly those used in critical healthcare applications during the ongoing pandemic.

Dataset Description

In this study, we used a publicly available X-ray repository called COVID-Net, available at GitHub (github.com/lindawangg/COVID-Net) [77], and a COVIDx-CT dataset that is a publicly accessible collection of CT images designed for COVID-19 research. Both these datasets have images of different COVID-19 pneumonia, non-COVID pneumonia, and normal patients.

There are various versions of COVID-NET available. We downloaded the COVIDx V9A dataset, which was updated on 26 November 2021. The dataset consists of more

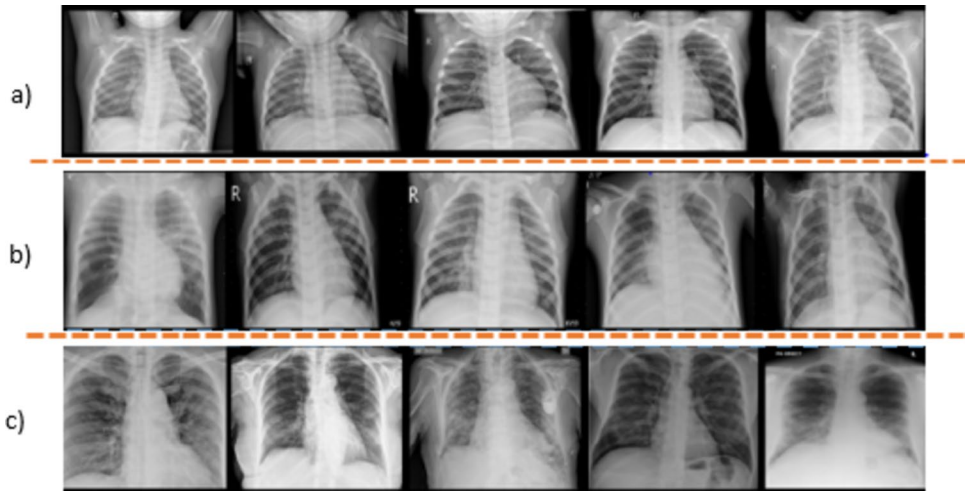

Fig. 2 Various types of chest X-ray images randomly selected from the dataset: **a** first row, normal (no pneumonia); **b** second row, non-COVID pneumonia; and **c** third row, COVID-19 pneumonia

than 30,000 CXR images for training, of which 16,490 are COVID-19-positive images from over 2800 patients, 5555 are non-COVID-19 pneumonia from over 5500 patients, and 8085 are normal images (no pneumonia) from over 8000 patients. Figure 2 shows the images from the dataset selected randomly from each class. There are 400 images for testing the model (100 normal, 100 pneumonia, and 200 COVID-19). The images in the training dataset are of a different dimension. However, we resized all the images to 224*224*3 during the preprocessing phase. During the attack, we randomly picked a bunch of images from the test data, computed the perturbation, and then evaluated the model's performance on imperceptible modified images. Table 1 summarizes the data distribution of X-ray images among different classes, and Table 2 shows the distribution of the number of patients in each category.

Similarly, we used the COVIDx CT-1 dataset on Kaggle, released on December 3, 2020. The dataset encompasses 104,009 CT slices derived from 1489 distinct patients. Derived from various open-source datasets, COVIDx-CT is continuously updated and enriched to enhance its utility and scope. COVIDx-CT dataset is released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license, in alignment with the licenses of its constituent datasets. It is important to note that certain subsets of the data may have less restrictive licenses, offering more flexibility in their usage. Figure 3 shows the images from the COVIDx CT-1 dataset selected randomly from each class. Table 3 shows the data distribution of CT images in the dataset.

Methodology

The work proposed in this study is visually represented in Figs. 4 and 5. Figure 4 outlines the comprehensive procedure of the adversarial attacking methodology. It displays two distinct paths labeled Path 1 and Path 2. Path 1 illustrates the standard training process of our specialized COVID-19 models, which are independently trained on X-ray and CT images. The COVIDx V9A dataset, a publicly available resource, was utilized for the X-ray images, while the CT images were sourced from COVIDx CT-1. This path involves a stage of data augmentation and preprocessing before the training of the models, while as Path 2, on the other hand, portrays the process of initiating an adversarial attack on the trained models. This phase includes preprocessing of the data and a dedicated adversarial attack module. We employed three separate adversarial example generators—FGSM, PGD, and BIM—to evaluate their performance thoroughly. We found that the models' overall performance, represented by the average accuracy, dropped significantly when challenged with adversarial examples. Figure 5 illustrates the proposed two-phase framework for robust classification against adversarial attacks. It underscores our use of the proposed adversarial learning algorithm during the training phase and JPEG transformation during the inference phase. This approach ensures the model's robustness and reliability during the training and inferencing phases, enhancing its resilience against adversarial attacks.

Table 1 COVIDx V9A chest radiography image distribution

Type	No pneumonia (normal)	Non- COVID pneumonia	COVID-19 pneumonia	Total
Train	8085	5555	16,490	30,130
Test	100	100	200	400

Table 2 COVIDx V9A patient distribution

Type	No. pneumonia (normal)	Non- COVID pneumonia	COVID-19 Pneumonia	Total
Train	8085	5531	2808	16,224
Test	100	100	178	378

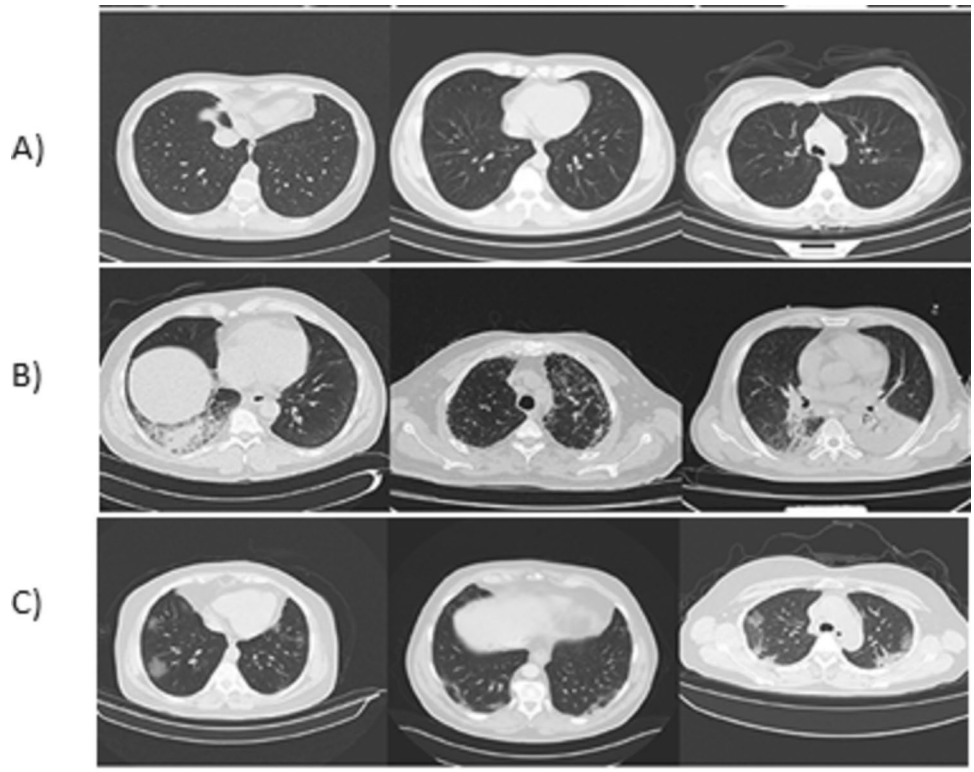

Fig. 3 Sample CT images from the COVIDx CT-1 benchmark datasets, representing different types of infections. **A** CT images of normal controls. **B** Features CT images of common pneumonia (CP). **C** CT images of novel coronavirus pneumonia (NCP) caused by SARS-CoV-2 infection. These images provide visual examples of the distinct characteristics associated with each infection type within the dataset

Platform and Data Preprocessing

This section provides a comprehensive overview of the platform details, including libraries and the data augmentation and preprocessing approach. We utilized the services of Google Colab to train, attack, and defend our deep COVID diagnosis model.

Platform Details

The development and experimentation of our deep learning models were conducted within a Python-based environment, leveraging a range of libraries and frameworks. We conducted our experiments within a Python 3.7 environment, utilizing TensorFlow and Keras for developing, attacking, and defending our deep learning models. We also employed sci-kit-learn for evaluation metrics calculation, ensuring a comprehensive assessment of our models' performance. Matplotlib was used for data visualization, while NumPy played a vital role in efficiently handling multidimensional arrays, especially during image preprocessing.

Table 3 CT Chest radiography image distribution

Type	No. pneumonia (normal)	Non- COVID pneumonia	COVID-19 pneumonia	Total
Train	3000	3000	3000	9000
Test	300	300	400	1000

This environment gave us the tools to effectively construct, evaluate, and analyze our deep learning models.

Data Augmentation and Preprocessing

Data augmentation and preprocessing are pivotal for preparing our dataset and addressing class imbalance issues. To enhance the representation of minority classes, specifically "no pneumonia (normal)" and "non-COVID Pneumonia," we employed data augmentation techniques. These involved generating augmented images by applying random transformations like rotation, flips, zoom, and brightness adjustments. This augmented the minority class data, mitigating imbalance and improving the model's overall performance. We harnessed TensorFlow's ImageDataGenerator for these augmentations, diversifying our training data effectively. Subsequently, we resized images to a uniform 224×224 pixel dimension using the same tool and normalized pixel values to a range between 0 and 1 with scikit-learn's MinMaxScaler. Label encoding was achieved seamlessly with TensorFlow's ImageDataGenerator class. These preprocessing steps primed our dataset for robust COVID-19 diagnosis model development.

Transfer Learning of Deep COVID-19 Diagnosis Model

We used the pre-trained state-of-the-art ResNet-50, VGG16, and inception-V3 as a base model and built our

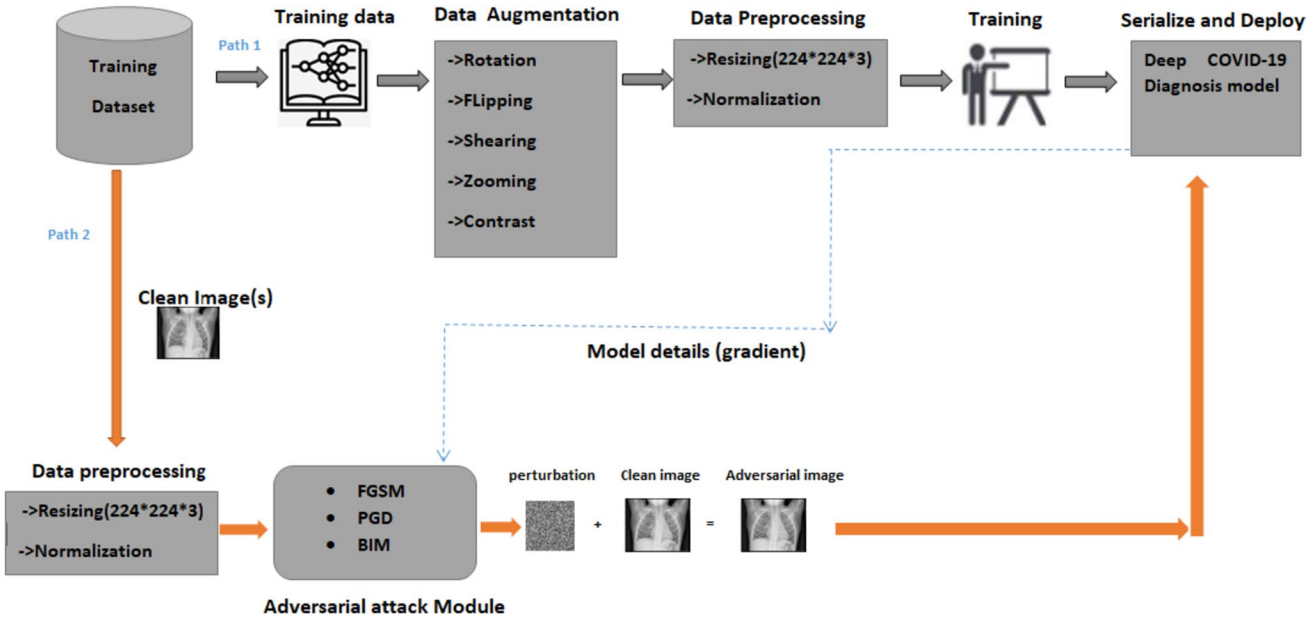

Fig. 4 The overall process of training a deep COVID-19 diagnosis model and attacking it with adversarial attacks. The blue dotted line indicates that the model details, such as gradients, are required to generate adversarial examples since the FGSM, PGD, and BIM are white-box attacks

COVID models by employing a transfer learning approach. We chose all these models because they are openly available and easy to modify. In particular, we selected them based on their performance in several image recognition projects. During the training of our models, we kept the layers of the base models frozen to preserve the learning of the weights that have already been learned. We removed the last layer in each of these base models and added 4 new

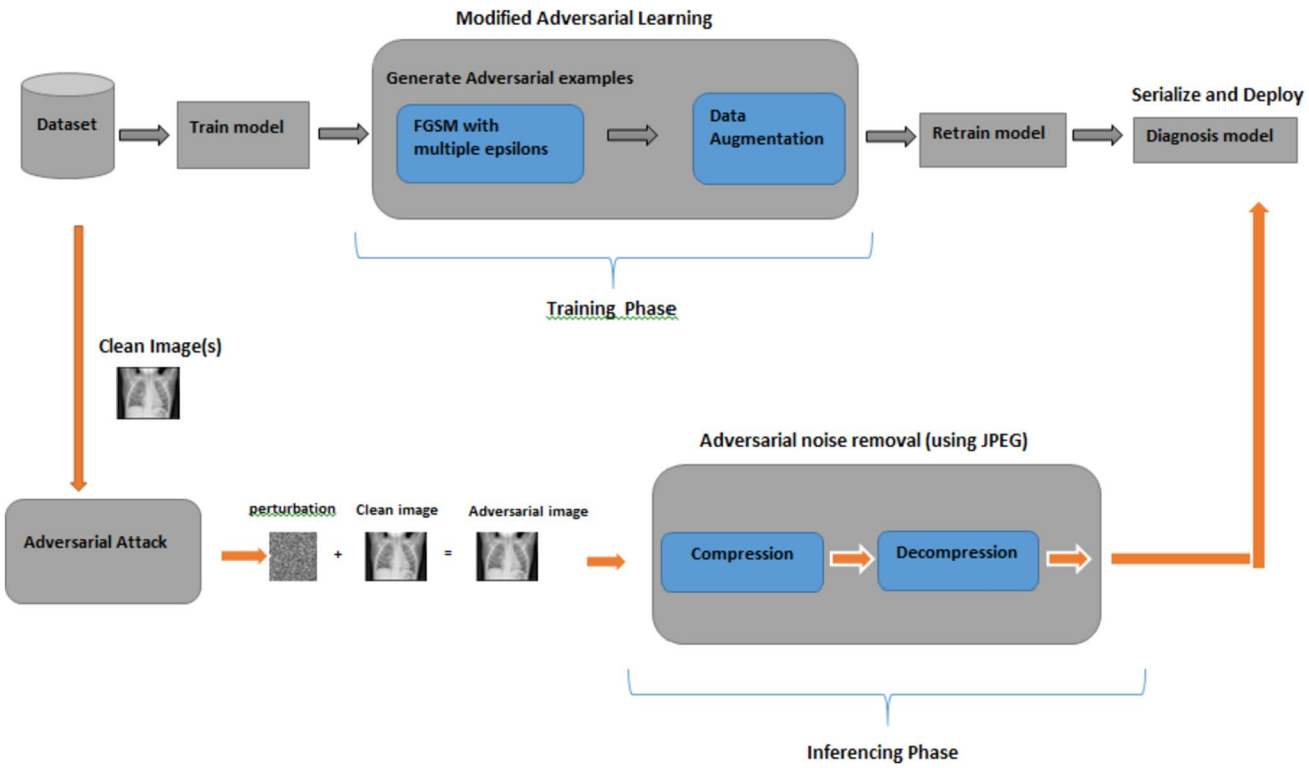

Fig. 5 Proposed two-phase framework for robust classification against adversarial attacks

trainable layers. Few dropouts and non-trainable pooling layers were added between these newly added layers to take care of overfitting and dimension reduction, respectively.

ResNet-50 Transfer Learning Model

The model is built on the ResNet-50 architecture, which contains 50 layers and utilizes skip connections or shortcuts to jump over some layers. This helps to solve the vanishing gradient problem and allows the model to learn deeper representations. By freezing the original layers and adding custom layers (Dense 64, Dense 32, and Dense 3), the model is tailored for COVID-19 detection, improving its efficiency and accuracy. The new layers are trained to adapt to the task, while the pre-trained layers provide a wealth of pre-existing features.

VGG-16 Transfer Learning Model

Built upon the VGG-16 architecture, this model consists of 16 layers, including 13 convolutional layers, successfully capturing spatial information and textual content in images. The remaining layers of the architecture are fully connected layers. By freezing the VGG-16 layers, the model retains its already learned weights. Custom layers (Dense 32 and Dense 3) are added to tune the model for the specific task of COVID-19 detection.

Inception-V3 Transfer Learning Model

This model is based on the Inception-V3 architecture, renowned for its computational efficiency. The architecture incorporates multilevel feature extraction through parallel convolutions, enabling the model to learn richer features. The original layers are frozen during training, preserving the learned weights. Custom layers (Dense 128, Dense 64, and Dense 3) are then appended to fine-tune the model for COVID-19 detection.

Adversarial Attack Preliminaries and Methodology

This section will first discuss several fundamental terminologies pertaining to attacks on deep learning-based models used for COVID-19 diagnosis.

- White-box attack: In a white-box attack scenario, the attacker possesses complete access to the targeted deep COVID-19 diagnosis model. This includes knowledge of the model's assets, such as gradients, training data, parameters, and architecture.
- Black-box attack: In a black-box attack scenario, the attacker does not have access to the internal workings of the deep COVID-19 diagnosis model. Instead, the model is publicly accessible through an application programming interface (API) that only allows input queries.
- Targeted attack: A targeted attack aims to manipulate the deep COVID-19 diagnosis model into producing a specific output label. For instance, the attacker may endeavor to force the model to classify any given chest X-ray image as COVID-19 positive.
- Untargeted attack approach: Unlike targeted attacks, untargeted attacks focus on misclassifying the deep COVID-19 diagnosis model without specifying a particular output label.
- Perturbation: Perturbation refers to introducing disturbance or noise into the original input to generate a perturbed input. This disturbance should be small enough to remain imperceptible to humans, yet significant enough to cause misclassification by the classifier.

The following symbols are as follows:
M = The deep COVID-19 diagnosis model.
x = The input image from the dataset.
y = The truth label of the input image.
Θ = Represents the parameters of the model.
J (θ, x, y) = The cost function used to train the neural network.
A = The step size.
δ = The adversarial perturbation added to the input image.
X_adv = The resulting adversarial example.
ε = The maximum perturbation allowed.
∇x J (θ, x, y) = computes the gradient of the loss function with respect to the input image.

Let us denote the deep COVID-19 diagnosis model as M. The model takes an input image x and outputs a probability distribution over different classes, including COVID-19, normal, and non-COVID pneumonia conditions. The objective function for the adversarial attack is defined as follows:

$$Min\|X_{adv} - x\|X \tag{1}$$

Subject to

$$M(x) = y$$

$$M\left(x^{'}\right)! = y$$

In the objective function, $M(x)$ correctly classifies the input, while $M(x')$ misclassifies the input as a different class. The objective function emphasizes that the magnitude of the perturbations introduced in the adversarial example should be small, but still significant enough to cause the model to misclassify the input..

The FGSM adds a small perturbation to the input image, enough to change the model's prediction. The perturbation is generated by calculating the gradient of the loss function with respect to the input image. The gradient points in the direction of the steepest ascent. Then, a small step in this direction is taken in the input space. Equation (1) is for FGSM:

$$\delta = \varepsilon * sign(\nabla x(\varnothing, x, y)) \tag{2}$$

$$X_{adv} = X + \delta$$

Here, ∇x J(θ, x, y) computes the gradient of the loss function with respect to the input image. The sign function creates a new perturbed image in the direction that will maximize the loss. ε (epsilon) is the magnitude of the perturbation, i.e., it bounds the total number of pixels in X_{adv} that can be modified with respect to x. The adversarial image X_{adv} is generated by adding the perturbation to the original image. This process aims to maximize the loss, leading to the model's misclassification of the adversarial image.

In contrast to FGSM, a one-step attack, PGD, is an iterative method, meaning that it applies FGSM multiple times with a small step size. A projection operation follows this to ensure that the adversarial example stays within the ε-ball of the original example in the pixel space. This iterative approach enables PGD to create more potent adversarial examples compared to FGSM. The following steps define the PGD adversarial attack:

1. Initialize x_adv = x
2. For each iteration, update x_adv by:

$$x_{adv} = P[x, \varepsilon]\left(x' + a * sign\left(\theta, x', y\right)\right) \tag{3}$$

where P represents the projection operation ensuring that x' lies within the ε-ball around x and the $P[x, \varepsilon]$ function ensures that x' stays within the ε-ball of x in the pixel space. This method tries to maximize the loss and thus aims to cause a misclassification, while the adversarial example remains visually indistinguishable from the original image.

BIM is an iterative variant of PGD where the step size is smaller. Like PGD, it uses multiple steps of size α to update the image, but it does not include the projection step. Instead, it clips the pixel values of the adversarial example after each update to ensure they stay in the ε-ball around the original image. The clipping operation is a simpler form of projection used in PGD. BIM applies the FGSM attack iteratively, and the adversarial example generation is given by: The BIM adversarial attack is defined by the following steps:

1. 1. Initialize x_adv= x.
2. For each iteration, update x' by:

$$x_{adv} = Clip[x, \varepsilon]\left(x' + a * sign\left(\nabla x' j\left(\theta, x', y\right)\right)\right) \tag{4}$$

The $Clip[x, \varepsilon]$ operation ensures that x' stays within the ε-ball of x in the pixel space by limiting (or clipping) the pixel values of x' to be within the range $[x—\varepsilon, x+\varepsilon]$. Like other adversarial attacks, BIM aims to maximize the loss, leading to a misclassification while ensuring the adversarial image remains visually similar to the original image.

In Fig. 6, we comprehensively illustrate the entire process involved in generating adversarial perturbations and examples. Notably, the adversarial image appears visually indistinguishable from the original image, making it challenging for human observers to discern any differences. However, what is particularly striking is that, despite this visual similarity, the model fails to classify the adversarial input correctly. This phenomenon underscores the insidious nature of adversarial attacks, where imperceptible alterations to input data can lead to significant misclassifications by even highly accurate models.

The Rationale for Using FGSM, PGD, and BIM

The incorporation of the fast gradient sign method (FGSM), projected gradient descent (PGD), and basic iterative method (BIM) in our study is underpinned by a host of advantages and a well-defined rationale. These advantages and rationales elucidate the methodological choices made in our research and underscore the significance of these adversarial attack techniques in the context of deep COVID-19 diagnosis models.

- Methodological Diversity
  One of the paramount advantages of employing FGSM, PGD, and BIM lies in their methodological diversity. FGSM represents a straightforward one-step adversarial attack, while PGD and BIM introduce iterative complexities. This diversity allows for a comprehensive evaluation of deep learning models across a spectrum of adversarial scenarios, mirroring real-world threat landscapes.
- Realism and Clinical Relevance
  FGSM, PGD, and BIM align with the realism and clinical relevance sought in our study. These methods emulate practical adversarial situations that medical diagnosis models may encounter in real-world healthcare settings. We gain insights into their robustness and vulnerabilities in scenarios that closely mimic clinical practice by subjecting the models to such attacks.
- Comparative Analysis
  Utilizing multiple adversarial attack techniques facilitates a comparative analysis of their effectiveness and impact. This comparative approach helps delineate the

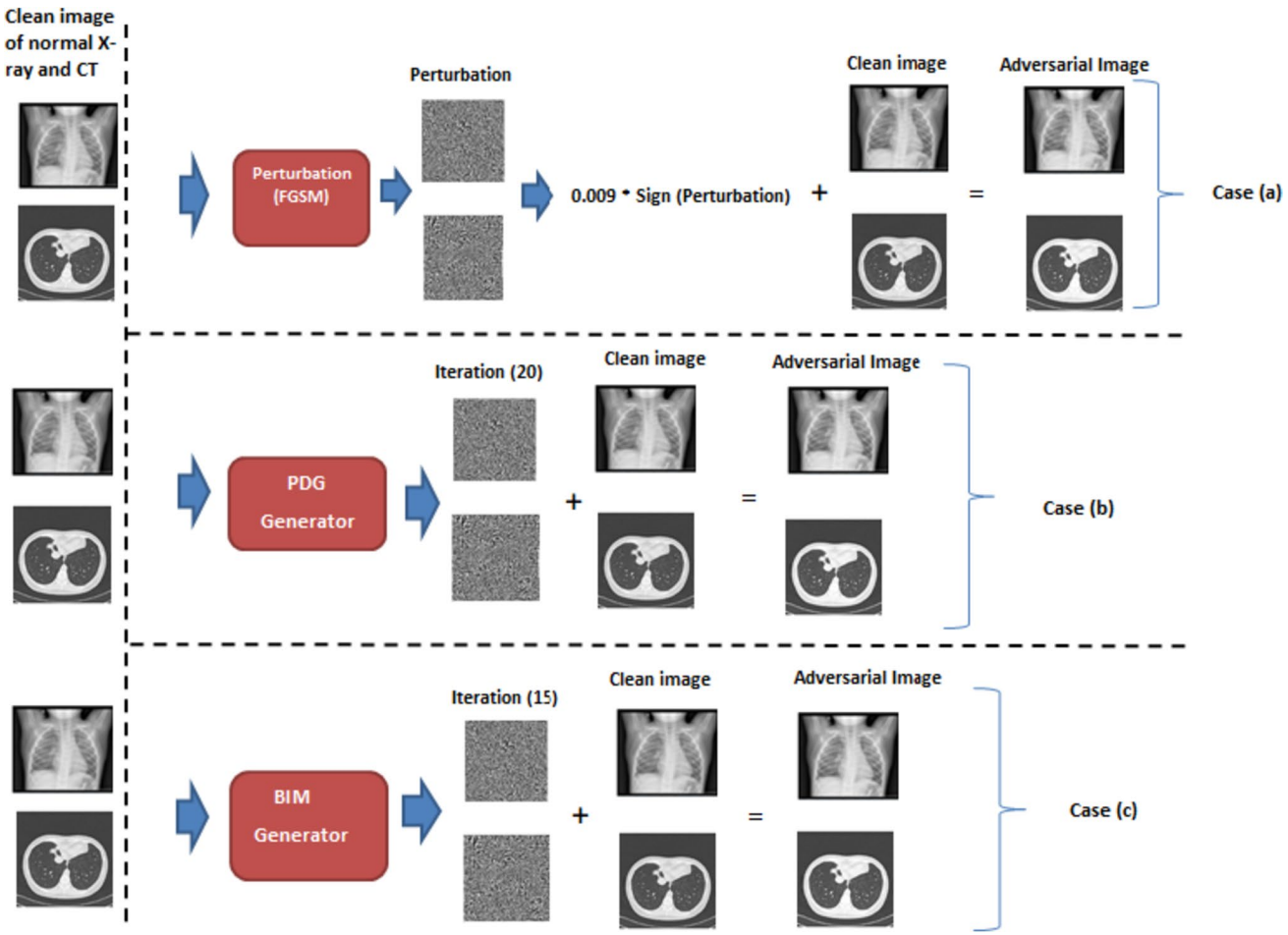

Fig. 6 Generation process of adversarial example from clean image: **a** process using FGSM technique by setting epsilon value to 0.009, **b** adversarial example generation via PGD attack, and **c** adversarial example using BIM approach

strengths and weaknesses of FGSM, PGD, and BIM, enabling us to understand their nuances. Such insights are invaluable for devising effective mitigation strategies and the development of robust countermeasures.

- Comprehensive Robustness Assessment

  FGSM, PGD, and BIM encompass a broad spectrum of adversarial complexities, ranging from simple, one-step attacks to more intricate iterative strategies. This comprehensive approach empowers us to assess the robustness of deep COVID-19 diagnosis models under varying levels of adversarial intensity and duration. Consequently, we obtain a holistic view of model performance across diverse adversarial conditions.
- Catalyst for Defense Strategies

  The deployment of FGSM, PGD, and BIM extends beyond mere vulnerability assessment. These adversarial attacks serve as catalysts for the development of robust defense mechanisms. Insights derived from these analyses inform the creation of effective defense strategies and adversarial training techniques, reinforcing the security of COVID-19 diagnosis models in clinical applications.
- Research Advancement

  Finally, the rationale for using FGSM, PGD, and BIM stems from their role in advancing the frontiers of adversarial research in medical deep learning. By exploring these techniques in the context of COVID-19 diagnosis, our study contributes to the broader field of adversarial machine learning, fostering the development of more secure and reliable medical AI systems.

Proposed Two-Phase Robust Framework

Our approach leverages modified adversarial learning during training and applies JPEG compression as a preprocessing step during inference. By training on diverse adversarial examples and employing compression to remove adversarial noise, we improved the model's resilience to adversarial attacks.

Modified Adversarial Learning During Training

Our proposed framework presents a robust approach based on the conventional adversarial learning paradigm. In this framework, we introduce an adversarial generator module that generates adversarial examples for a given image at different epsilon values, surpassing the limitations of the single epsilon value approach. Additionally, we incorporate data augmentation techniques, including rotation, scaling, and contrast adjustments, to enhance the diversity of the generated adversarial images.

The primary objective of this framework is to improve the model's resilience against a broad range of adversarial perturbations, rather than focusing solely on a single perturbation. By considering multiple epsilon values and applying data augmentation, we aim to create a robust model capable of accurately classifying instances, even in the presence of sophisticated adversarial attacks. To illustrate the implementation of our approach, Algorithm 1 provides a step-by-step outline of the modified adversarial learning process.

Step 1 of the algorithm is to generate adversarial examples. The algorithm generates adversarial examples for each image in the regular dataset for each epsilon value in set E. Hence, the time complexity of this step is $O(N * |E|)$, where N is the number of images in the regular dataset. The algorithm computes the loss between the predicted label and the true label for the clean image to generate the adversarial example. Then, it computes the gradient of the loss with respect to the input image and generates the adversarial example by adding the perturbation to the clean image. The algorithm applies data augmentation to the generated adversarial examples to increase the diversity of the adversarial examples. Finally, the adversarial example is added to the modified training dataset. In STEP 2, during the training process, the model is optimized using a weighted combination of the original loss and adversarial loss, which encourages the model to correctly classify the clean images and the adversarial examples generated in Step 1. This process is repeated for T iterations, which allows the model to learn from the modified dataset and improve its robustness continuously. The gradient of the total loss with respect to the model parameters is computed, and the model parameters are updated using stochastic gradient descent with a learning rate α. Therefore, the time complexity of this step is $O(T * B)$, where T is the number of iterations and B is the batch size. In Step 3, the algorithm evaluates the robust model by computing its accuracy on a test set of images. The time complexity of this step depends on the size of the test set, which we will denote as M. Hence, the time complexity is $O(M)$. In Step 4, it is required to improve the robustness of the model and avoid overfitting. Repeating steps 1–3 with a different random seed can generate a different set of adversarial examples and further diversify the training data. This helps the model to learn a more generalized decision boundary that can better distinguish between clean and adversarial examples, thereby improving its robustness. The number of repetitions will be denoted as R. Therefore, the time complexity of this step is $O(R * (N * |E| + T * B + M))$. Overall, this algorithm improves the robustness of a non-robust model to adversarial attacks by generating a diverse set of adversarial examples and using them to modify the training dataset.

The overall time complexity of algorithm 3 is approximately $O(R * (N * |E| + T * B + M))$, where R is the number of repetitions, N is the number of images in the regular dataset, $|E|$ is the number of epsilon values, T is the number of iterations, B is the batch size, and M is the size of the test set. This complexity analysis accounts for the time required to generate adversarial examples, train the robust model, evaluate its accuracy, and repeat the process with different random seeds to enhance dataset diversity. Additional factors such as model architecture, optimization algorithm efficiency, and hardware specifications may influence the actual runtime.

JPEG Compression Preprocessing During Inferencing

In the inference phase, we introduce a preprocessing module that applies JPEG compression to the test data samples. JPEG compression is a widely used lossy image compression technique that reduces file size by removing unnecessary details. However, it also has the effect of removing high-frequency noise, including adversarial perturbations. By applying JPEG compression to the test images before passing them to the model, we effectively remove adversarial noise, enabling the model to make accurate classifications.

Here are the steps to use JPEG compression for adversarial noise removal:

1. Generate adversarial images: First, an adversarial attack is performed on the original images to create adversarial examples.
2. Apply JPEG compression: Next, apply JPEG compression on the adversarial images. This step involves transforming the image to a different space (discrete cosine transform (DCT)), quantization, and encoding. During the quantization step, high-frequency components that contain adversarial perturbations are discarded, resulting in an image with reduced adversarial noise.
3. Decompress images: After compression, the images are decompressed back to their original size. During this process, the adversarial noise originally added to the image is reduced or even eliminated.

To illustrate this, let us represent the adversarial image as A, the JPEG compression function as C(·), and the

Input: regular dataset, label, non-robust model, E (set of epsilon values), number of iterations T, weight factor λ, learning rate α, adversarial image (x_adv), Adversarial example set (x_adv_set)

Output: Robust model

STEP 1: Generate adversarial examples

STEP 1.1: x_adv_set = [ ]
For *each image in the regular dataset*, **do**:
For *each ε in the E*, **do**:

STEP1.2: Compute the loss between the predicted label and the true label:
loss = CategoricalCrossEntropy(label, non-robust model.predict(clean image)).

STEP1.3: Compute the gradient of the loss with respect to the input image:
gradient = gradient(loss, clean image).

STEP1.4: Generate the adversarial example by adding the perturbation to the clean image:
x_adv = clean image + ε * sign(gradient).

STEP 1.5: Apply data augmentation to the adversarial example to increase diversity:
x_adv= apply_random_augmentation(x_adv)

STEP1.7: Add the adversarial example to the modified training dataset:
x_adv_set.append (x_adv)

STEP 2: **Train the robust model**

For *T iterations*, **do** :

STEP2.1: Sample a batch of images from the modified training dataset

STEP2.1: Compute the loss between the predicted label and the true label:
loss = CategoricalCrossEntropy(label, robust model.predict(batch))

STEP2.1: Compute the adversarial loss between the predicted label and the true label for the adversarial examples in the batch:
adversarial_loss = CategoricalCrossEntropy(label, robust model.predict(adversarial batch))

STEP2.1: Compute the total loss as a weighted sum of the original loss and the adversarial loss:
total_loss = loss + λ * adversarial_loss

STEP2.1: Compute the gradient of the total loss with respect to the model parameters:
gradient = gradient(total_loss, model parameters)

STEP2.1: Update the model parameters using stochastic gradient descent with learning rate α:
model parameters = model parameters - α * gradient

Step 3: Evaluate the robust model

compute the accuracy of the robust model on a test set of images

Step 4: Repeat steps 1-3 with a different random seed

To increase the diversity of the modified training dataset and avoid overfitting, repeat steps 1-3 with a different random seed.
Take the average accuracy over all runs as the final accuracy.

Algorithm 1 Modified adversarial learning

decompression function as D(·). The following steps can represent the process:

A = Original Image + Adversarial pertubation

$$A_{compressed} = \mathrm{C(A)}$$

$$A_{decompressed} = \mathrm{D}(A_{compressed})$$

After these steps, $A_{decompressed}$ is the final image where adversarial noise has been significantly reduced. This image can then be input into the model for classification.

The proposed algorithm 2 requires several input parameters to be provided before its execution. The first parameter is the adversarial image (*x_adv*), representing the input image subjected to adversarial perturbations. The next parameter is the JPEG compression quality factor (*q*), which determines the JPEG compression quality applied to the adversarial image. The quality factor is typically specified as a numerical value ranging from 0 to 100. Higher values (e.g., 90) indicate higher image quality with less compression, while lower values (e.g., 10) indicate lower image quality with more compression. The choice of the quality factor depends on the desired trade-off between image compression and the preservation of important image details.

The threshold for adversarial detection (*t*) is another important parameter. This threshold determines whether an image is still considered adversarial after JPEG compression. It is usually set as a value between 0 and 1. A lower threshold value (e.g., 0.1) indicates a stricter criterion for considering an image as adversarial, while a higher threshold value (e.g., 0.5) indicates a more lenient criterion. The selection of the threshold depends on the desired robustness level and the application's specific requirements.

The maximum number of iterations (*max_iter*) is a parameter that sets the upper limit on the number of iterations allowed in the optimization process to remove adversarial noise. It is typically set as a positive integer value. A higher value allows for more iterations, which may result in better noise reduction but also increases the computational time. The choice of the maximum number of iterations depends on factors such as the complexity of the adversarial noise and the available computational resources. The convergence threshold (epsilon) is the final parameter, determining the criterion for stopping the iteration loop. It is typically set as a small positive value, such as 0.01 or 0.001. A smaller epsilon value indicates a stricter convergence criterion, requiring more precise noise reduction before stopping the iterations. The choice of the convergence threshold depends on the desired level of noise reduction and the trade-off with computational resources.

The algorithm involves several steps to process the adversarial image and remove adversarial noise using JPEG compression. The pretrained deep learning model *M* is initially loaded into memory (Step 1), serving as the foundation for subsequent operations. The specific adversarial image *x_adv* is then loaded as the input for the algorithm (Step 2), acting as the starting point for the noise

Input: Adversarial image (x_adv), Set JPEG compression quality factor (q), threshold for adversarial detection (t), number of iterations (max_iter), convergence threshold (epsilon)
Output: Clean Image (x_reconstructed)

Step 1: Load the pre-trained deep learning model M
Step 2: Load the adversarial image x_adv
Step 3: Set iteration counter iter = 0
Step 4: Set delta = x_adv
Step 5: Compress x_adv using JPEG compression with quality factor q to obtain x_compressed
Step 6: Decompress x_compressed to reconstruct x_reconstructed
Step 7: Repeat until iter reaches max_iter or convergence is achieved:
  Step 7.1: Pass x_reconstructed through M to obtain predicted class probabilities y_pred
  Step 7.2: Compute adversarial confidence as max(y_pred) - y_pred[target_class]
  Step 7.3: If adversarial confidence < t, exit loop
  Step 7.4: Compute gradient of loss function with respect to x_reconstructed: ∇_x J (M(x_reconstructed), target_class)
  Step 7.5: Update delta = delta - epsilon * sign (∇_x J(M(x_reconstructed), target_class))
  Step 7.6: Clip delta to ensure it remains within a valid range
  Step 7.7: Update x_reconstructed = x_adv + delta
  Step 7.8: Increment iter by 1
Step 8: Output final reconstructed image x_reconstructed.

Algorithm 2 JPEG preprocessing to remove adversarial noise

removal process. The iteration counter *iter* is initialized to zero (Step 3), allowing the algorithm to keep track of the number of iterations performed. The optimization process begins by setting the variable delta to the adversarial image *x_adv* (Step 4), representing the adversarial noise that will be gradually reduced. JPEG compression is applied to the adversarial image using a specified quality factor *q* (Step 5), effectively reducing the file size while preserving important visual information. The compressed image is then decompressed to reconstruct the image (Step 6), aiming to restore the original image while mitigating adversarial noise. The algorithm enters an iterative optimization loop (Step 7), continuing until convergence or reaching the maximum number of iterations. In each iteration, the reconstructed image is passed through the deep learning model M to obtain predicted class probabilities (Step 7.1), allowing evaluation of the model's response. The algorithm computes the adversarial confidence by comparing the maximum predicted probability with the predicted probability of the target class (Step 7.2), quantifying the level of adversarial perturbation present. The algorithm exits the loop if the computed confidence falls below the specified threshold (Step 7.3). If the adversarial noise is still significant, the gradient of the loss function with respect to the reconstructed image is computed (Step 7.4), providing information about the direction and magnitude of the loss change in response to variations in the image. The delta is updated by subtracting the product of a learning rate and the sign of the gradient (Step 7.5), modifying the adversarial noise to reduce its impact. To ensure valid values, the updated delta is clipped to an appropriate range (Step 7.6). Finally, the reconstructed image is updated by adding the delta, effectively integrating the modifications made to the adversarial noise (Step 7.7).

The algorithm continues to iterate through the optimization loop, encompassing the steps of passing the reconstructed image through the model, computing adversarial confidence, updating the delta, and clipping the delta until convergence is achieved or the maximum number of iterations is reached. The time complexity primarily depends on the size of the image for JPEG compression and decompression, typically *O(N log N)*, where *N* represents the image size. The time complexity of the iterative optimization loop depends on the model architecture and can be approximated as *O(N)* or *O(N log N)*.

Experimental Results

We utilized the services of Google collaborator to train, attack, and defend our deep COVID diagnosis model. The model and the attack were implemented in a Python 3.7 environment using Tensorflow and Keras packages. We represent the Resnet-50-based model as (M1), VGG-16 as (M2), and inceptionV3 as (M3). The models were validated on 400 test samples. We evaluated the performance of the models (performance on clean data, adversarial attacks and performance on proposed defense approach) separately on the X-ray and CT datasets.

Evaluation Metric

1. **Accuracy (ACC)**

   - Accuracy is a fundamental metric that measures the overall correctness of predictions made by a model.
   - It is defined as the ratio of correctly predicted instances to the total instances:

     $$ACC = (TP + TN)/(TP + TN + FP + FN)$$

     Where:

     – TP (true positives) represents correctly predicted positive instances.
     – TN (true negatives) represents correctly predicted negative instances.
     – FP (false positives) represents instances predicted as positive but are negative.
     – FN (false negatives) represents instances predicted as negative but are positive

Table 4 Hyperparameter setting of a deep COVID-19 diagnosis model during training

Hyperparameter	ResNet-50 (M1)	VGG-16 (M2)	Inception-V3 (M3)
Input layer_size	224 *224*3	224 *224*3	224 *224*3(modified)
Batch_size	32	32	32
Epochs	50	50	50
Learning_rate	1e-4 with decay rate = learning rate/ epoch number;	1e-3 with decay rate = learning rate/ epoch number;	1e-3 with decay rate = learning rate/epoch number;
Dropout_rate	0.3, 0.5, 0.4	0.5, 0.5, 0.5	0.5, 0.4, 0.5
Optimization	ADAM	ADAM	ADAM
Loss function	CategoricalCrossentropy	CategoricalCrossentropy	CategoricalCrossentropy
Output layer_size	224*224*3	224*224*3	224*224*3 (modified)

Table 5 Parameters used in the adversarial attacks

Attack	Parameters
FGSM	$\varepsilon = 0.005$
PGD	$\varepsilon = 0.2$, $\alpha = 2/255$, steps = 20
BIM	$\varepsilon = 0.04$, $\alpha = 1/255$, steps = 15

2. **Precision (P)**
   - Precision quantifies the accuracy of positive predictions made by a model.
   - It is defined as the ratio of true positives to the total predicted positives:

$$P = TP/(TP + FP)$$

3. **Recall (sensitivity, true-positive rate)**
   - Recall measures the model's ability to identify all relevant instances, particularly the positive ones.
   - It is defined as the ratio of true positives to the total actual positives:

$$\text{Recall} = TP/(TP + FN)$$

4. **F1-score**
   - The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics:

$$F1 - \text{Score} = (2 \cdot P \cdot \text{Recall})/(P + \text{Recall})$$

5. **Specificity (true-negative rate or selectivity)**
   - It measures the ability of a model to correctly identify negative instances (the "true negatives") out of all actual negative instances.

$$\text{Specificity} = TN/(TN + FP)$$

Experimental Setup

During the training of the models, the parameter setting that was used is shown in Table 4. The parameters used during the adversarial attack are shown in Table 5. The parameters used during the proposed defense approach are shown in Tables 6 and 7.

Evaluation of X-ray Dataset

We initially assessed the performance of our deep learning models (M1, M2, and M3) using clean X-ray data in the Pre-Attack Evaluation subsection. Following that, we thoroughly examined these models' resistance to adversarial attacks, including FGSM, PGD, and BIM, in the Post-Attack Evaluation subsection. Finally, in the Post-Defense Evaluation subsection, we demonstrated the robustness of our framework against these attacks, highlighting its effectiveness in enhancing model performance.

Pre-attack Evaluation on X-ray

The models were extensively trained for 50 epochs on the X-ray dataset, encompassing three distinct classes: "normal," "pneumonia," and "COVID-19." Table 8 presents a comprehensive overview of the performance metrics for three distinct models, M1, M2, and M3, on an X-ray dataset. These metrics, including average accuracy, precision, recall, F1-score, and specificity, provide a robust assessment of model performance over multiple iterations. For example, Model M1 consistently maintained an impressive average accuracy of 93.34%, affirming its capability to distinguish between normal, pneumonia, and COVID-19 cases. Model M2, closely following, demonstrated an average accuracy of 91.45%, signifying its proficiency in precisely identifying the three classes. Notably, Model M3 exhibited exceptional performance, achieving the highest average accuracy of 94.65% among the models. This underlines its superiority in classifying X-ray images effectively. The average accuracy of the models on testing data over multiple epochs is shown in Fig. 7.

Furthermore, we have visually presented the model outputs for all three X-ray images, namely "normal," "non-COVID pneumonia," and "COVID-19," in Fig. 8. A meticulous examination of the figure reveals the exemplary performance of the models. For instance, Model M1 exhibited remarkable accuracy by correctly classifying a clean "normal" X-ray image as "normal" patient with an impressive confidence level of 99.73%.

Post-Attack Evaluation on X-ray

In "COVID-19 Diagnosis Models", "Contribution", Introduction, Table 8 revealed how well the models performed

Table 6 Parameters used during the training phase in the proposed defense approach

Parameter	Value	Description
Epsilon (ε)	0.001, 0.005, 0.009	The reason for producing adversarial images with varying epsilon values is to ensure that the model is resilient to a variety of adversarial perturbations instead of a single perturbation
Data augmentation	Rotation, contrast	Increase the diversity of the adversarial examples
Epochs (T)	50	This means that the optimization process will run for a maximum of 100 iterations

Table 7 Parameters used during the inferencing phase in the proposed defense approach

Parameter	Value	Description
Quality factor (q)	80	This represents a moderate compression level, striking a balance between image quality and file size reduction
Threshold for adversarial detection (t)	0.2	Indicating that an adversarial image is successfully mitigated if the adversarial confidence falls below 0.2
Max_iter	100	This means that the optimization process will run for 100 iterations
Convergence epsilon (epsilon)	0.001	This indicates that the algorithm will stop iterating once the changes in the adversarial noise reduce below this threshold

when analyzing X-ray images under normal conditions. They achieved high accuracy, f1-score, and specificity, indicating their ability to correctly classify these images into different categories. However, when we subjected these models to adversarial attacks, the situation changed significantly, as shown in Table 9. For instance, with the FGSM attack, M1's accuracy dropped to 18.37%, M2's to 17.29%, and M3's to 18.59%. The PGD attack was even more challenging, resulting in accuracy rates of just 9.17% for M1, 8.31% for M2, and 8.74% for M3. Finally, the BIM attack had a similar impact, with M1 achieving an accuracy of 8.68%, M2 at 7.15%, and M3 at 9.84%.

Figure 9 presents a comprehensive analysis of each model's performance under adversarial attacks, highlighting their vulnerability to such malicious inputs as evidenced by a substantial drop in accuracy. Furthermore, the impact of adversarial attacks on model outputs is depicted in Fig. 10. For example, an "adversarial-normal-image" generated using FGSM is alarmingly misclassified by M1 as a COVID-19-infected person with a staggering 99.33% confidence level, whereas in Fig. 8, M1 correctly identifies the clean version of the same image as "normal' with 99.73% confidence. The concern lies in M1's high-confidence misclassification of the adversarial normal image, emphasizing the crucial need for robust defenses and countermeasures to bolster the models' resilience against such attacks.

Post-Defense Evaluation on X-ray

The primary aim of our proposed defense framework is to enhance a model's ability to classify adversarial

Table 8 Accuracy, precision, recall, F1-score, and specificity of the models on clean X-ray test data

	M1				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Specificity (%)**	**Support**
	M1				
No pneumonia (normal)	92.49	96.47	94.55	93.42	100
Non-COVID pneumonia	96.31	90.31	93.27	92.47	100
COVID_pneumonia	94.48	94.37	94.56	92.82	200
Average accuracy	93.34				400
Macro_average	93.12	92.56	92.41	92.56	400
Weighted_average	93.34	92.37	92.93	92.82	400
	M2				
No pneumonia (normal)	92.73	96.26	93.28	90.31	100
Non-COVID pneumonia	93.75	90.38	91.57	89.38	100
COVID _pneumonia	92.58	91.41	94.93	91.93	200
Average accuracy	91.45				400
Macro_average	91.41	90.49	90.49	90.89	400
Weighted_average	90.76	91.58	91.41	91.08	400
	M3				
No pneumonia (normal)	92.73	97.65	95.45	94.49	100
Non-COVID pneumonia	98.93	92.37	92.47	94.85	100
COVID _pneumonia	94.47	93.77	95.67	94.62	200
Average accuracy	94.65				400
Macro_average	95.57	94.28	94.52	94.51	400
Weighted_average	94.48	94.24	94.42	94.67	400

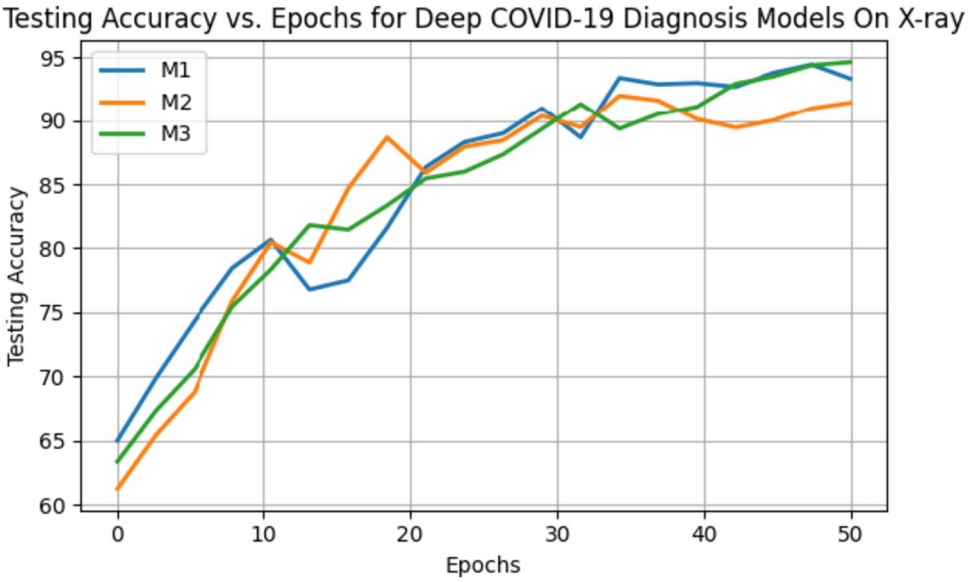

Fig. 7 Testing accuracy of the models over 50 epochs

images. Nevertheless, we also comprehensively evaluated the framework's performance on clean inputs. This assessment sought to gauge the effectiveness of the defense mechanism in enhancing the model's performance when faced with various input scenarios(adversarial and clean samples).

In Table 10, we provided the performance of the post-defense models on the clean X-ray images. The accuracy, precision, recall, and specificity percentages for models M1, M2, and M3, trained on the X-Ray dataset, are reported. The models achieved a slight drop in average accuracies on the clean inputs, with M1 achieving 92.89%, M2 achieving

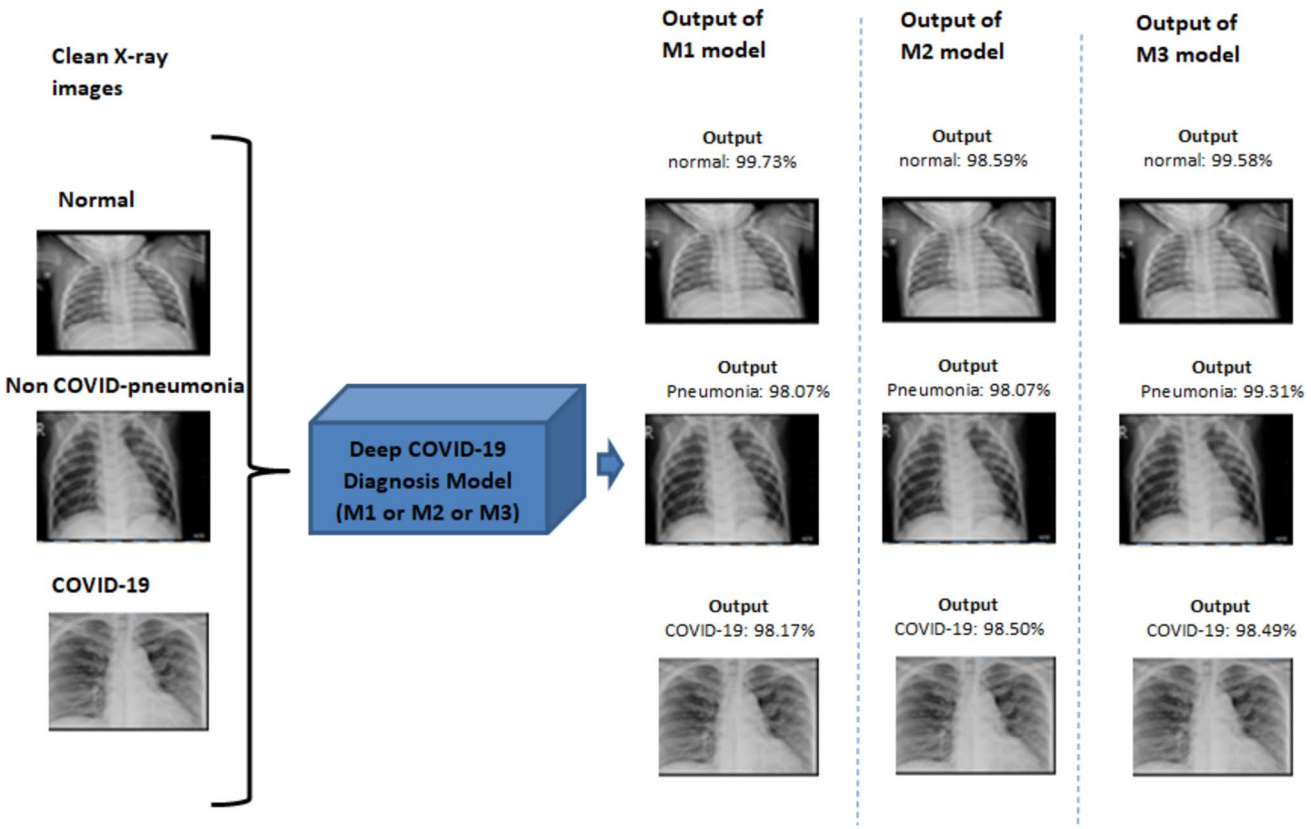

Fig. 8 Models performed exceptionally well on X-ray images, predicting correctly with a very high confidence score

Table 9 Average accuracy, F1-score, and specificity of the models on adversarial X-ray images

Attacks	M1 (%)			M2 (%)			M3 (%)		
	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity
FGSM	18.37	18.05	18.56	17.29	17.15	16.06	18.59	17.31	16.48
PGD	9.17	9.56	9.03	8.31	7.79	7.20	8.74	7.92	6.89
BIM	8.68	8.23	8.97	7.15	7.17	7.97	9.84	8.73	8.16

91.23%, and M3 achieving 94.28%. These values indicate the models' ability to accurately classify the clean X-ray images after implementing the defense mechanism. Figure 11 compares pre-defense and post-defense models regarding accuracy on a clean X-ray dataset.

In Table 11, we showcase the post-defense models' performance on various adversarial attacks in terms of average accuracy, F1-score, and specificity. Despite the presence of adversarial inputs, the models maintain relatively high performance. For instance, under the FGSM attack, M1 achieves an average accuracy of 94.39%, M2 achieves 92.82%, and M3 achieves 93.79%. Similarly, the PGD and BIM attacks yield accuracies that demonstrate the models' resilience, with M2 exhibiting an accuracy of 93.04%, M3 achieving 92.89% under the PGD attack, M1 achieving an accuracy of 92.85%, and M3 achieving 94.73% under the BIM attack. The performances of post-defense models on adversarial attacks in terms of average accuracies are shown in Fig. 12.

The results affirm the effectiveness of our proposed defense framework in enhancing the model's performance on both clean and adversarial inputs. This defense mechanism enables the accurate classification of clean X-ray images and provides robustness against adversarial attacks. Figure 13 illustrates the robust model's performance on adversarial inputs, where it excels. For example, an "adversarial normal X-ray" image was correctly classified as normal with 98.52% confidence. In contrast, the non-robust model in Fig. 10 misclassified the same adversarial image as COVID-19 with 99.3% confidence. These findings emphasize our defense mechanism's practicality and real-world effectiveness, highlighting its potential to enhance the reliability and security of machine learning models in healthcare applications.

Evaluation of CT Dataset

Likewise, we assessed the models' performance using high-resolution CT images. While CT images are relatively

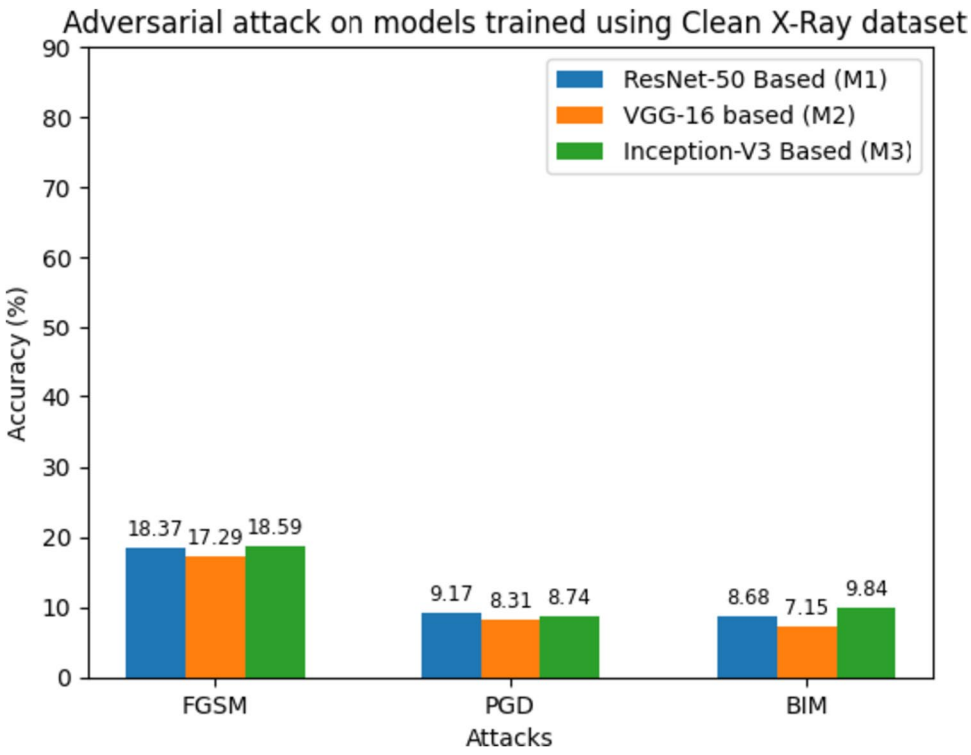

Fig. 9 Pre-defense (non-robust) models' average accuracy on each adversarial attack

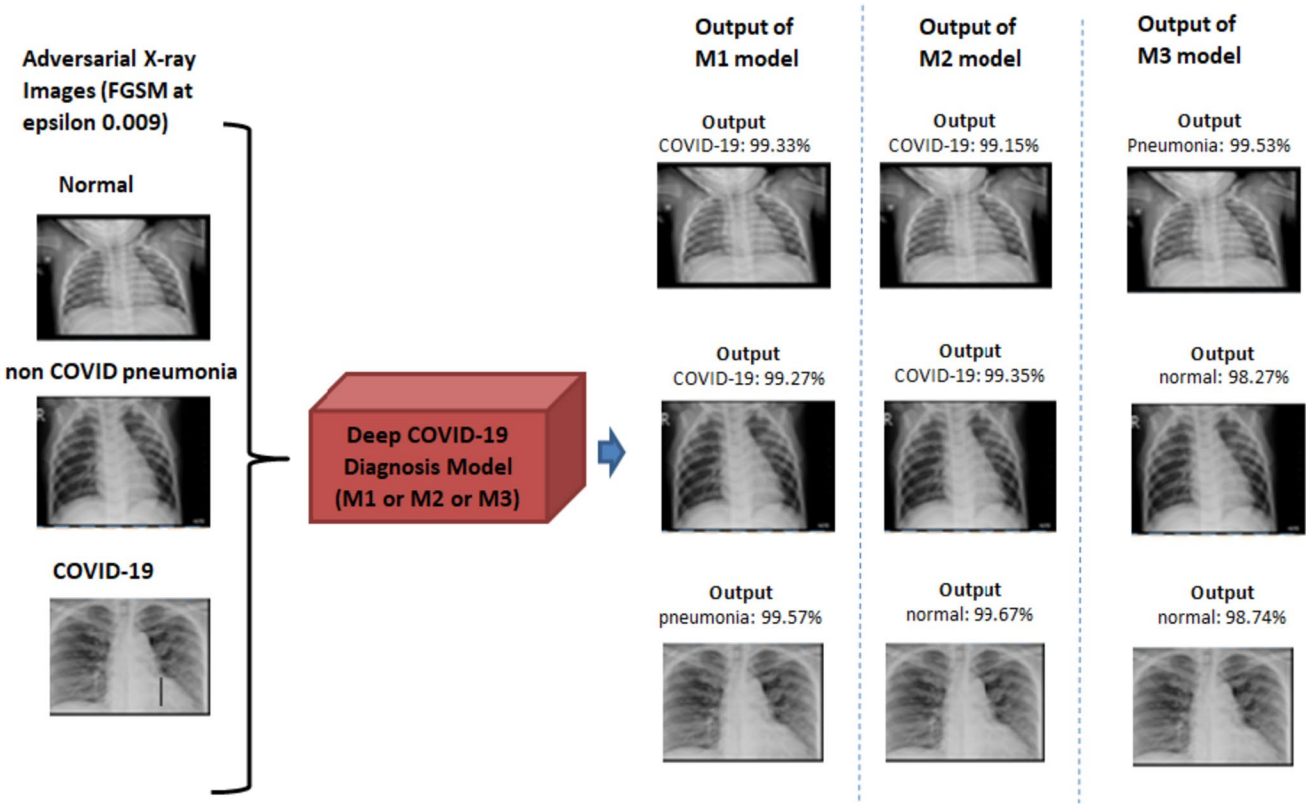

Fig. 10 Models' output on adversarial examples along with their confidence scores. It is apparent from the figure that all the models consistently misclassified the adversarial images with remarkably high confidence levels

Table 10 Performance of post-defense models (after implementing the proposed defense approach) on the clean X-ray images

	M1				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Specificity (%)**	**support**
	M1				
No pneumonia (normal)	92.29	96.37	94.25	93.03	100
Non-COVID pneumonia	96.24	89.89	93.14	92.25	100
COVID _pneumonia	93.23	94.23	94.43	92.23	200
Average accuracy	**92.89**				400
Macro_average	92.32	92.23	92.29	92.23	400
Weighted_average	92.79	92.89	92.34	92.76	400
	M2				
No pneumonia (normal)	92.32	97.31	93.03	90.21	100
Non-COVID pneumonia	96.43	90.29	91.27	89.67	100
COVID _pneumonia	92.87	92.25	95.78	91.89	200
Average accuracy	**91.23**				400
Macro_average	91.41	90.49	90.34	90.32	400
Weighted_average	90.76	91.58	91.52	90.83	400
	M3				
No pneumonia (normal)	92.65	97.28	95.24	94.35	100
Non-COVID pneumonia	98.78	92.14	92.39	94.26	100
COVID _pneumonia	94.39	93.48	95.52	94.49	200
Average accuracy	**94.28**				400
Macro_average	95.34	94.21	94.34	94.32	400
Weighted_average	94.21	94.09	94.19	94.09	400

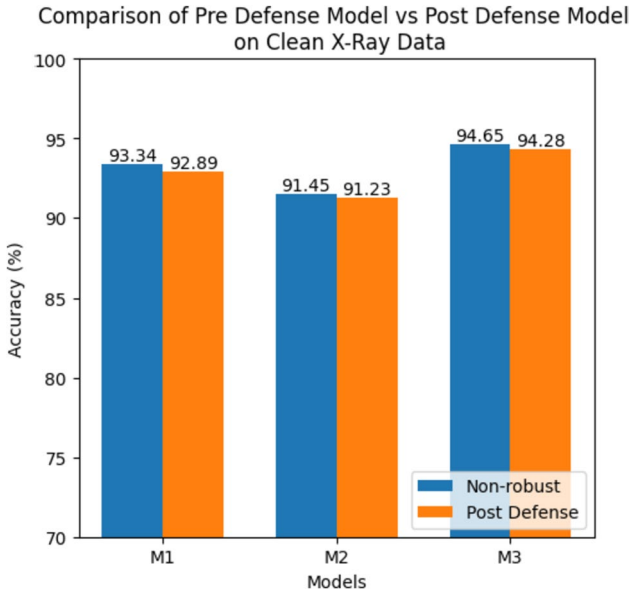

Fig. 11 Comparison of pre-defense models (before implementing the proposed defense approach) and post-defense models (after implementing the proposed defense approach) on a clean X-ray dataset

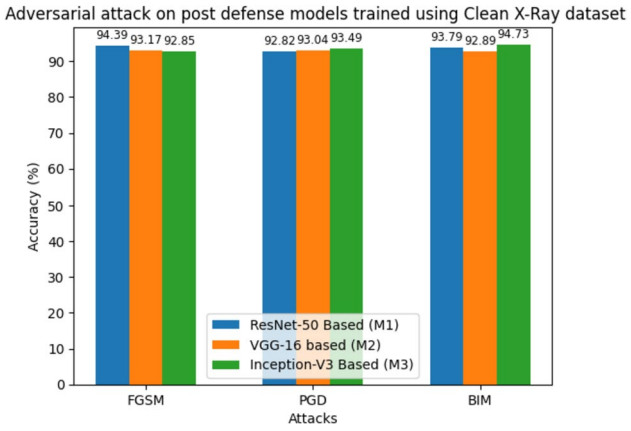

Fig. 12 Performance of the post-defense (robust) models on adversarial examples

costly and less prevalent than X-ray images, this evaluation aimed to validate the applicability of our proposed defense approach across various image modalities.

Pre-attack Evaluation on CT

Similarly, Table 12 illustrates the performance metrics of the models (M1, M2, and M3) on the clean CT dataset. The average accuracies for the models are as follows: M1 achieves an accuracy of 95.72%, M2 achieves 92.89%, and M3 achieves 95.03% on the clean CT dataset. Model M1 achieved the highest accuracy of 98.21% on the CT dataset. In Fig. 14, we presented the testing accuracy curve to represent how the accuracy evolved across 50 epochs visually. Figure 15 demonstrates how the model classifies the clean CT image.

Post-Attack Evaluation on CT

The performance drastically reduced when the same models were exposed to multiple adversarial attacks, as shown in Table 13. It focuses on the accuracy of the models trained on the clean CT dataset when subjected to adversarial attacks. The attacks considered are FGSM, PGD, and BIM. The results indicate a significant drop in accuracy compared to the clean dataset. For instance, under the FGSM attack, M1 achieves an accuracy of 21.82%, M2 achieves 20.02%, and M3 achieves 20.97%. Similarly, the PGD and BIM attacks result in lower accuracy percentages for all models. Figure 16 shows the comparison of the performance of the models on adversarial examples. Figure 17 demonstrates the classification capabilities of the models on adversarial examples generated via the FGSM approach. For instance, M1 misclassified the adversarial normal image as a COVID-19 patient with 99.99% confidence. The same version of the clean image was classified as normal with 99.73% confidence in Fig. 15.

Post-Defense Evaluation on CT

After implementing the proposed framework, we evaluated the model's performance on both clean and adversarial inputs separately. The reason was to check the proposed defense work's impact on both clean and adversarial inputs. Table 14 presents the performance of the post-defense models on the clean CT dataset. After implementing the proposed defense mechanism, the models' accuracies remain

Table 11 Performance of post-defense models (after implementing the proposed defense approach) on the adversarial examples generated by FGSM, PGD, and BIM

Attacks	M1 (%)			M2 (%)			M3 (%)		
	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity
FGSM	94.39	93.89	94.39	92.82	92.37	93.18	93.79	93.10	93.76
PGD	93.17	93.28	93.27	93.04	93.18	93.01	92.89	92.51	92.34
BIM	92.85	92.65	92.73	93.49	93.48	93.27	94.73	94.49	94.16

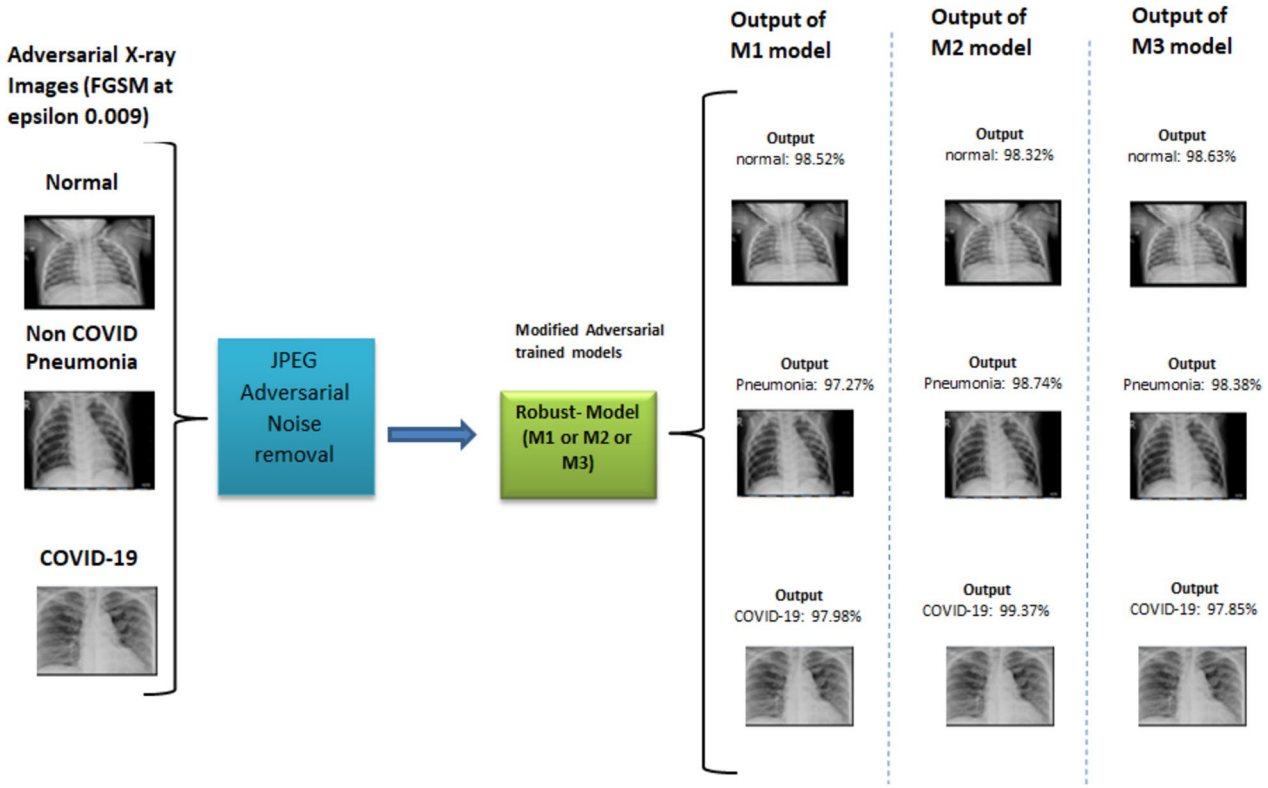

Fig. 13 Models' performance after implementing our proposed defense approach. The process begins by subjecting adversarial images to a preprocessing step involving a JPEG-based adversarial noise removal filter. Subsequently, these preprocessed images are fed into the models trained using the modified adversarial learning approach. The results are striking as models exhibit remarkable proficiency in classifying adversarial images and correctly identifying them confidently

relatively high. M1 achieves an accuracy of 94.29%, M2 achieves 94.19%, and M3 achieves 95.33% on the clean CT dataset post-defense. In addition, we have compared the accuracy of the pre-defense and post-defense models on clean and CT datasets in Fig. 18.

Table 15 focuses on the performance of the post-defense models when exposed to adversarial attacks using CT images. Notably, the accuracies of the models increase compared to Table 14, indicating the effectiveness of the defense mechanism. For instance, under the FGSM attack, M1 achieves an accuracy of 95.93%, M2 achieves 94.01%, and M3 achieves 94.95%. Similar trends are observed for the PGD and BIM attacks. Figure 19 shows the performance of the post-defense models on CT image datasets. Figure 20 shows the output of the robust models on CT images when exposed to adversarial images.

Ablation Study and Discussion

We conducted an ablation study to ascertain the individual contributions of adversarial learning and adversarial image filtering to the overall robustness of our proposed two-phase defense framework against adversarial attacks. This investigation allowed us to systematically evaluate the efficacy of the individual elements of our methodology.

Effect of Adversarial Learning Alone

We scrutinized the impact of adversarial learning without implementing adversarial image filtering for the first phase of our ablation study. Our advanced adversarial learning algorithm trained the models, while the adversarial image filtering stage during inference was omitted. This approach provided insight into the extent of resilience contributed by adversarial learning against adversarial attacks. The model's performance on modified adversarial learning is shown in Table 16. The findings from this analysis facilitated an understanding of the standalone strength of adversarial learning in enhancing the model's defense against these attacks.

Effect of Image Filtering Alone

The second phase of our ablation study sought to discern the independent influence of image filtering in the absence

Table 12 Performance of the models in terms of the precision recall, accuracy, F1-score, and specificity on clean CT test images

	M1				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Specificity (%)**	**Support**
			M1		
No pneumonia (normal)	93.05	96.85	95.39	96.15	100
Non-COVID pneumonia	96.93	91.48	95.43	93.74	100
COVID _pneumonia	94.81	94.85	94.57	95.56	200
Average accuracy	95.72				400
Macro_average	94.62	94.83	95.06	94.29	400
Weighted_average	94.37	94.96	94.85	94.85	400
			M2		
No pneumonia (normal)	93.38	96.85	93.89	92.79	100
Non-COVID pneumonia	96.96	90.85	91.73	93.25	100
COVID_pneumonia	92.89	92.78	95.60	92.25	200
Average accuracy	92.89				400
Macro_average	92.73	92.52	92.04	92.89	400
Weighted_average	92.38	92.81	92.39	92.72	400
			M3		
No pneumonia (normal)	93.85	97.65	96.66	95.38	100
Non-COVID pneumonia	97.60	94.37	94.85	95.31	100
COVID_pneumonia	97.83	95.20	95.98	94.49	200
Average accuracy	95.03				400
Macro_average	95.40	94.37	94.94	94.35	400
Weighted_average	94.75	94.86	94.31	94.57	400

of adversarial learning. The models were conventionally trained in this setup, foregoing the adversarial learning step. However, adversarial image filtering through JPEG compression was applied during the inference phase. This experiment allowed us to gauge the performance of the image filtering technique in safeguarding a model not specifically trained to withstand adversarial attacks. The results are shown in Table 17.

The findings from this ablation study are instrumental in unveiling the individual effectiveness of adversarial learning and image filtering in the face of adversarial attacks. This will aid us in enhancing the individual components

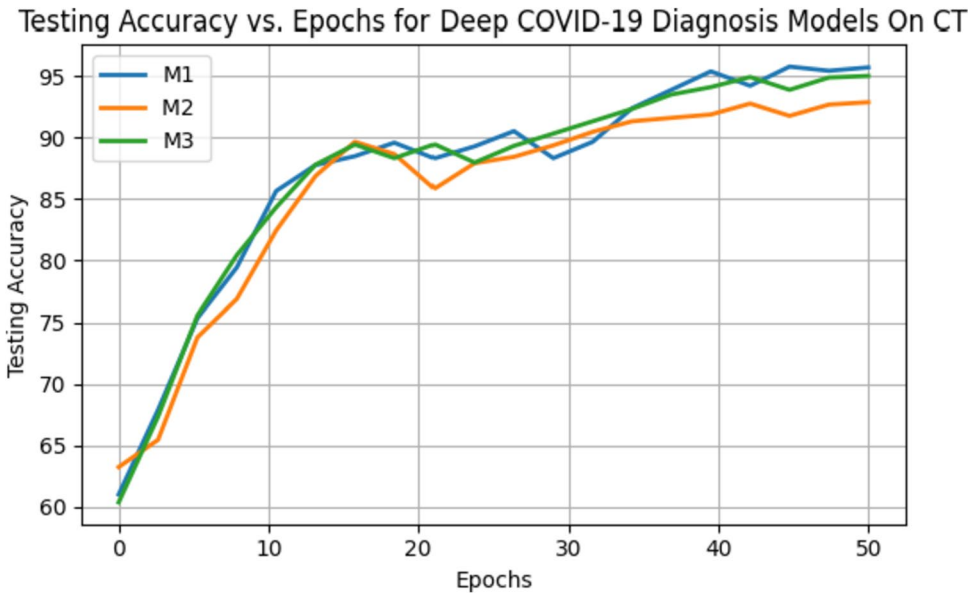

Fig. 14 Testing accuracy curve of all three models on unseen samples over 50 epochs

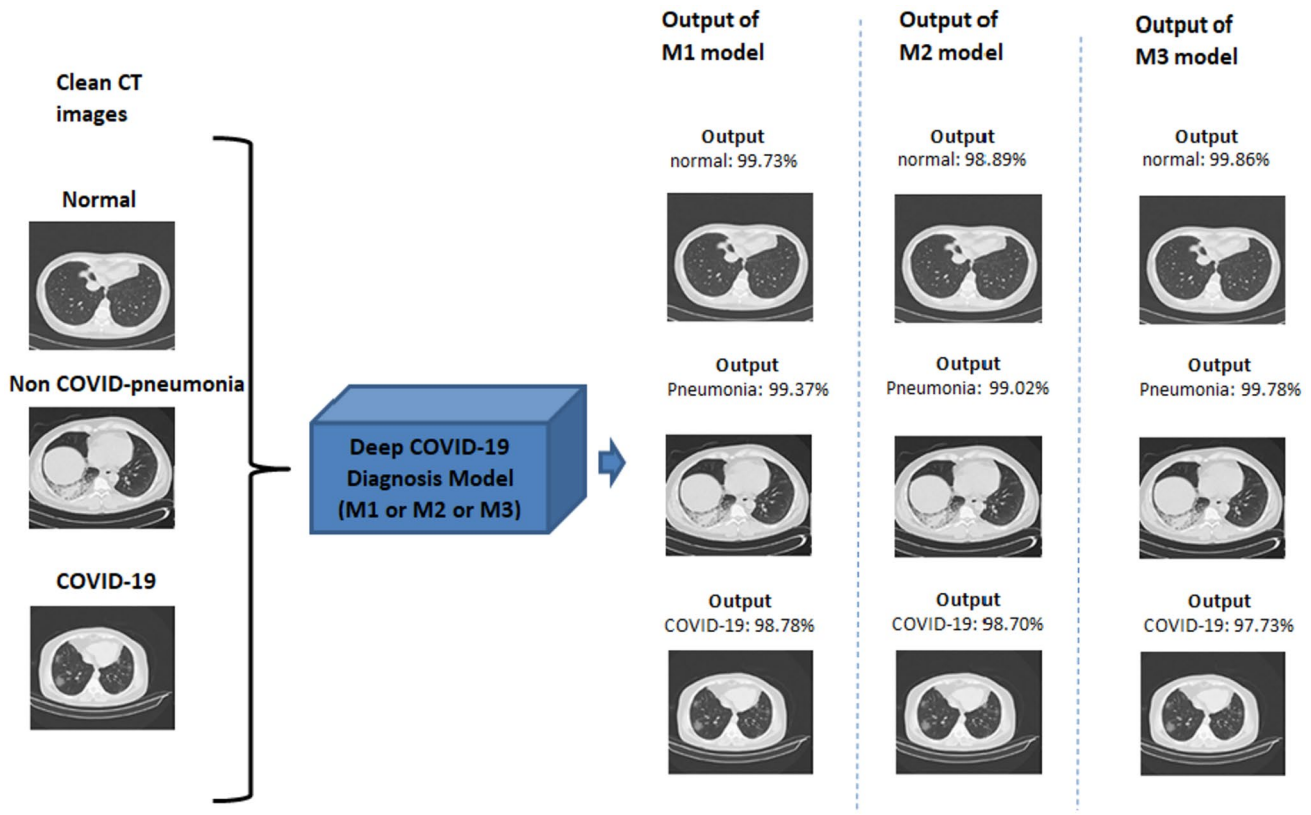

Fig. 15 Remarkable performance of the models in accurately classifying clean CT images with high confidence scores

and the integrated system to ensure high reliability and accuracy in AI-based COVID-19 diagnosis. We hope to publish these results in forthcoming reports for further scrutiny and application.

Comparison of the Developed Models Against State-of-the-Art Models

First, we compare the developed COVID-19 diagnosis models with state-of-the-art models in Table 18. It is evident that the proposed models outperform them in terms of average accuracy. Moreover, in Table 19, we compare the previous work of the COVID-19 battling tools regarding their vulnerabilities against adversarial attacks and the defense techniques to make them robust against attacks.

In terms of attack performance, the proposed method achieves an accuracy of over 92% under FGSM, PGD, and BIM adversarial attacks. This is notably higher compared to the other works, which report attack performances ranging from 60.82% to 91%. Regarding defense performance, the proposed method employs a two-phase security approach, resulting in a defense performance of over 95%. This defense performance exceeds that of other works, which report defense performances ranging from 80 to 90%. The higher defense performance indicates that the proposed method effectively mitigates the impact of adversarial attacks and provides robustness to the models.

Table 13 Average accuracy, F1-score, and specificity of the models on the adversarial attack

Attacks	M1 (%)			M2 (%)			M3 (%)		
	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity
FGSM	21.82	20.55	21.39	20.02	20.15	19.47	20.97	20.43	19.87
PGD	12.34	12.08	11.38	9.56	8.95	7.78	10.23	10.74	9.99
BIM	8.59	8.37	7.86	7.52	7.45	7.02	9.02	8.91	8.44

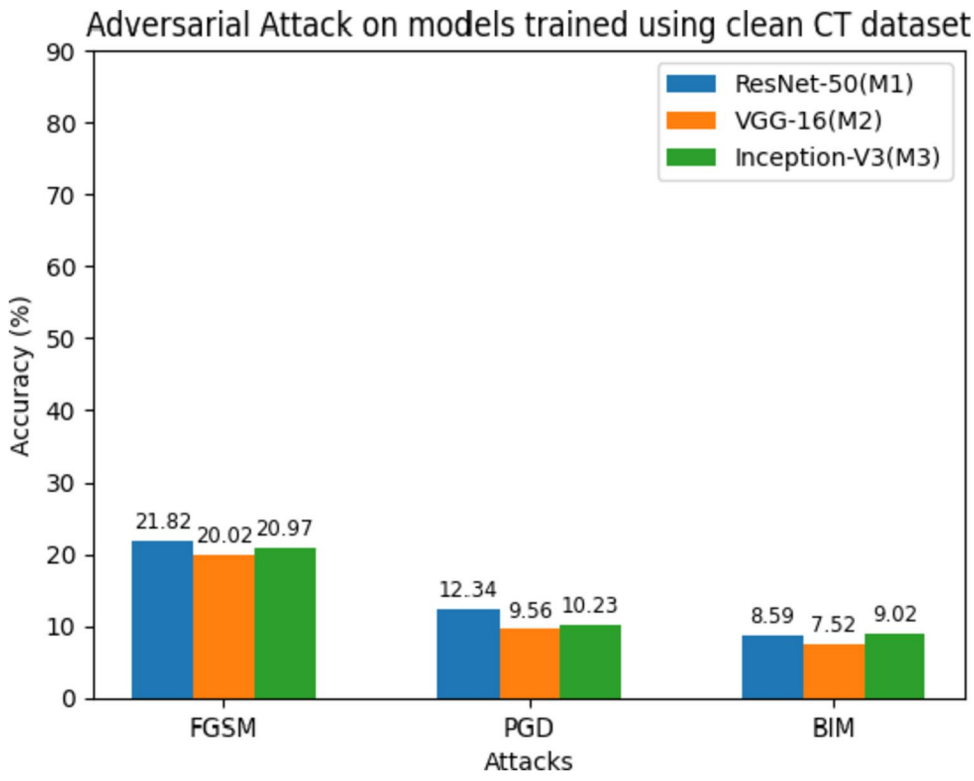

Fig. 16 Average accuracy of the pre-defense models after the adversarial attacks

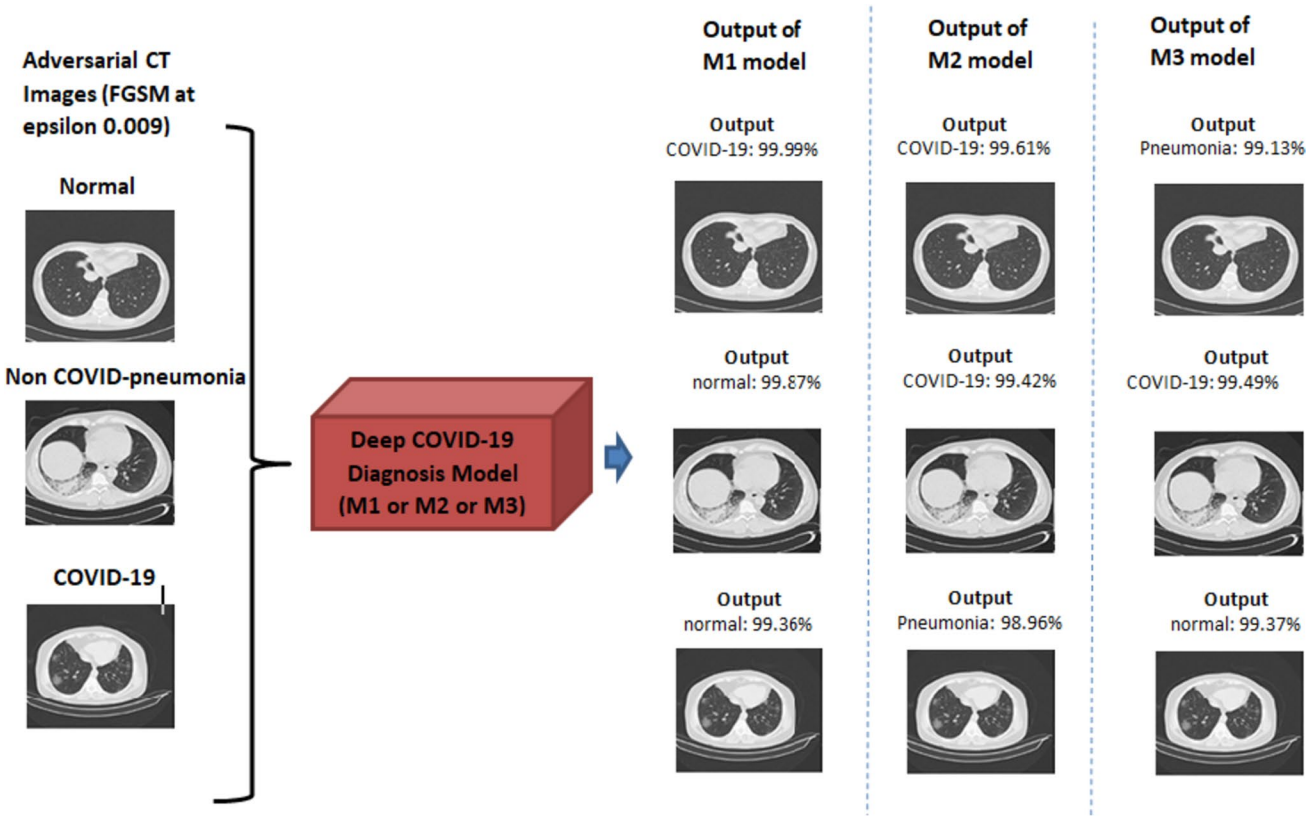

Fig. 17 Misclassification of the models on adversarial images with very high confidence values

Table 14 Performance of post-defense models on clean CT images

	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)	Support
	M1				
No pneumonia (normal)	93.26	95.67	94.86	95.67	100
Non-COVID pneumonia	95.89	91.23	95.18	93.06	100
COVID _pneumonia	94.17	94.01	94.63	95.17	200
Average accuracy	94.29				400
Macro_average	94.10	93.95	94.06	94.18	400
Weighted_average	94.07	94.31	94.27	94.31	400
	M2				
No pneumonia (normal)	93.89	96.92	94.62	93.05	100
Non-COVID pneumonia	97.47	91.23	92.34	93.75	100
COVID_pneumonia	93.46	92.91	95.89	92.67	200
Average accuracy	94.19				400
Macro_average	94.20	94.12	94.08	94.31	400
Weighted_average	94.33	93.97	93.88	94.18	400
	M3				
No pneumonia (normal)	93.91	97.88	96.82	95.83	100
Non-COVID pneumonia	97.73	94.68	95.05	95.58	100
COVID_pneumonia	97.80	95.62	96.10	94.73	200
Average accuracy	95.33				400
Macro_average	95.14	95.11	95.41	95.35	400
Weighted_average	94.65	95.24	94.76	94.89	400

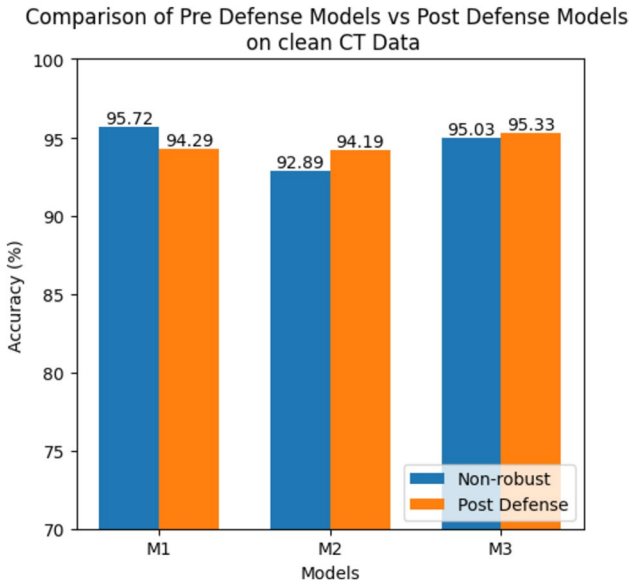

Fig. 18 Accuracy of the pre and post-defense models on clean CT dataset

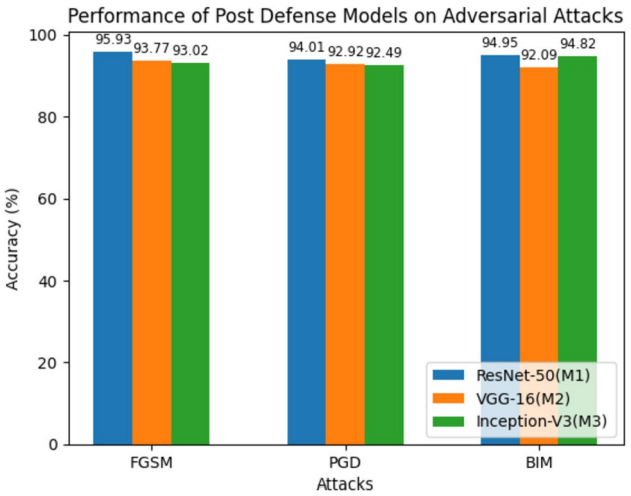

Fig. 19 Average accuracy of post-defense models that are trained using CT images on the adversarial examples

Table 15 Performance of post-defense models on adversarial attacks generated by FGSm, PGD, and BIM

Attacks	M1 (%)			M2 (%)			M3 (%)		
	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity	AVG ACC	AVG_F1-score	AVG_ Specificity
FGSM	95.93	94.22	95.16	94.01	94.87	94.14	94.95	93.53	93.26
PGD	93.77	93.37	93.48	92.92	93.18	93.06	92.09	92.34	92.53
BIM	93.02	93.21	93.33	92.49	92.20	92.41	94.82	93.84	94.48

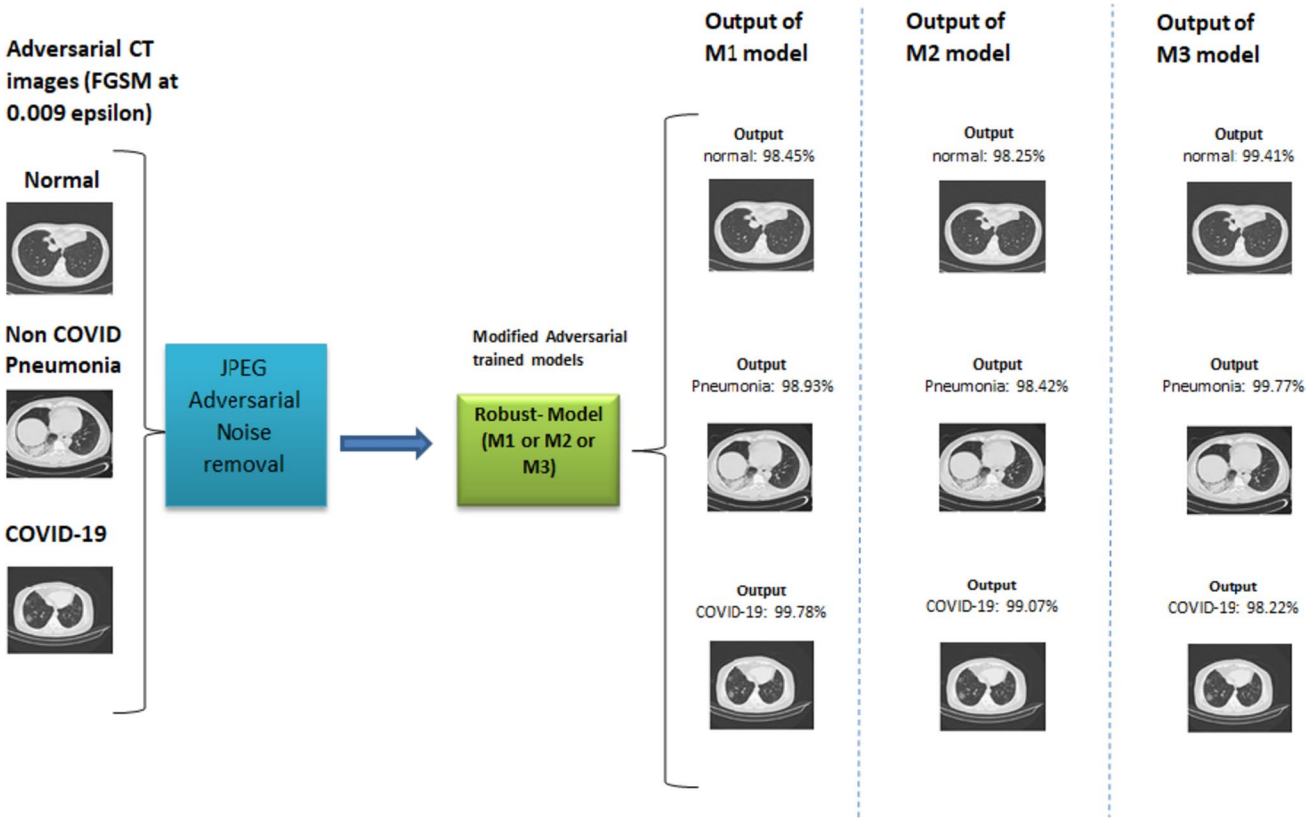

Fig. 20 Performance of the proposed robust framework on adversarial images. Each image was correctly classified with very high confidence. For instance, the robust model M1 successfully classifies a "normal adversarial image" as a "normal' patient with a 98.45% confidence score. However, the same image was misclassified by the M1 in Fig. 17 as COVID-19 with a 99.99% confidence score

Table 16 Performance of models on X-ray with only modified adversarial learning

Attacks	M1 (ACC %)	M2 (ACC %)	M3 (ACC %)
FGSM	85.24	83.45	86.29
PGD	81.01	80.40	82.09
BIM	84.92	80.18	82.97

Table 17 Performance of models on X-ray with only JPEG compression on adversarial attacks

Attacks	M1 (%)	M2 (%)	M3 (%)
FGSM	60.38	62.19	68.29
PGD	59.28	58.32	54.39
BIM	52.18	49.03	53.82

Table 18 Comparison of the accuracy of the diagnosis models on clean images against state-of-the-art models

Model	Average accuracy (trained on X-ray)	Average accuracy (trained on CT)
M1	93.34	95.72
M2	91.45	92.89
M3	94.65	95.03
[78]	87.02	
[79]	89.50	
[80]	92.85	
[81]	90.60	

Table 19 Comparison of adversarial attacks and defense strategies used in previous works

Article	Models	Dataset	Attack method	Attack performance	Defense method	Defense performance
[67]	COVID-Net [68]	X-ray	UAP	> 90%	Adversarial training	> 80%
[69]	ResNet,YOLO, DarkNet, GRAD-CAM	CT scan X-ray	FGSM,MIFGSM, DF, LBFGS, C&W, BIM, FB, PGD, JSMA, BD, MS, poisoning	91%	-	-
[71]	VGG, ResNet	CelebA		90% in white box setting and 80% in black box setting	-	-
[72]	VGG 16 InceptionV3	X-ray, CT scan	FGSM	83.3	-	-
[73]	U-Net	CT scan	Stabilized medical image attack (SMIA)	60.82%	-	-
[74]	ResNet-18	X-ray	FGSM PGD	94.7%	-	-
[75]	ResNet50, ResNet18, WRN-16–8, VGG-19, InceptionV3	X-ray	FGSM, PGD, C&W, ST	> 60%		
[76]	MobileNet-V2	Custom dataset	FGSM	> 83%	Adversarial learning	~ 90%
Proposed method	ResNet50, VGG-16, Incepton-V3-based model	COVIDx V9A and COVIDx CT-1	FGSM, PGD, BIM	> 92%	Two-phase security	> 95%

Conclusion

This research delves into the critical issue of adversarial attacks against deep learning models utilized in COVID-19 diagnosis. Our comprehensive exploration has led us to discover potential vulnerabilities of such models. Despite the inherent robustness and high performance of our developed deep learning-based COVID-19 diagnosis models using ResNet-50 (M1), VGG-16 (M2), and Inception-V3 (M3), they proved susceptible to FGSM, PGD, and BIM adversarial attacks. This vulnerability poses a significant risk, especially in a sensitive field like healthcare, where reliability and accuracy are paramount.

To address these vulnerabilities, we proposed a novel two-phase defense strategy. The application of this technique solidified the model's resilience against adversarial attacks, maintaining high accuracy even when exposed to adversarial examples.

Our experimental findings can be summarized as follows:

- Without adversarial attacks, the deep learning models demonstrated commendable accuracy, ranging from 91.45% to 94.65% on the X-ray dataset and 92.89% to 95.72% on the CT dataset.
- Upon exposure to adversarial attacks such as FGSM, PGD, and BIM, there was a significant deterioration in the performance of the models, with their accuracy plunging to as low as 7.15% to 18.59% on the X-ray dataset and 7.52% to 21.82% on the CT dataset.
- Following deploying our proposed two-phase defense framework, the models exhibited remarkable resilience against adversarial attacks. Their performance improved drastically, achieving an accuracy of over 92% against all three types of attacks across both datasets.
- Through an ablation study, we discerned the individual impact of adversarial learning and image filtering on enhancing the resilience of the models.
- Implementing adversarial learning alone improved the models' performance ranging from 80.18% to 86.29%, while applying image filtering alone increased performance from 49.03% to 68.29%. This finding indicates the complementary role of both techniques in bolstering the defense against adversarial attacks.

The results of our study have been encouraging. Our defense mechanism successfully mitigated the adverse effects of adversarial attacks, and the models retained high performance, making them more reliable for diagnosing COVID-19 from radiology images. However, the journey towards a more secure and robust AI in healthcare is still ongoing. We believe that the defense mechanism proposed in this study is a step forward. However, further research is required to explore and

uncover more comprehensive and efficient defense mechanisms, considering the rapidly evolving adversarial attack techniques. We believe that our findings will assist researchers in improving the security of their models and raise awareness of the need to establish deep COVID-19 diagnosis models with several protection strategies.

Limitations and Future Scope

The current study, despite its advancements and contributions, also has certain limitations:

- JPEG compression dependency: The defense framework relies heavily on using JPEG compression. While this has proven to be effective in the study, it might introduce some image degradation and information loss, which could impact diagnostic accuracy, especially when high-resolution imaging details are critical.
- Dataset constraints: The models are trained and evaluated using a specific dataset. Variations in imaging protocols, patient demographics, disease stages, and the quality of radiology images across different datasets might affect the model's performance and robustness against adversarial attacks.
- Scalability and generalizability: While the study showed promising results, it was only tested on three specific deep learning models (ResNet-50, VGG-16, and Inception-V3). How well the defense framework would scale or generalize to other architectures or DL models is unclear.
- Single modality: This study focuses exclusively on radiology images (X-ray and CT). Its effectiveness for other types of medical imaging modalities (e.g., MRI, PET) or multimodal diagnostic data has not been explored.
- Applicability beyond COVID-19: While the study has great relevance for COVID-19 diagnostics, its direct applicability to models designed for diagnosing other diseases or medical conditions remains untested.
- Computational requirements: Our two-phase defense framework, particularly the adversarial training phase, can be computationally intensive and time-consuming, which might limit its applicability in systems with constrained resources.

In the future, we would like to work on the following areas:

- Assessing other adversarial attacks: Future studies should assess the resilience of deep learning models against a wider array of adversarial attacks beyond FGSM, PGD, and BIM, such as the Carlini and Wagner attacks [82]. We also aim to test the robustness of the proposed model on other problem domains such as, robustness against adversarial attack in video surveillance and network intrusion detection systems [83–86].
- The proposed work explored the vulnerability in a white-box environment. However, in the future, we aim to investigate the model's performance in the black box, the environment in which the model's details are not visible to the attacker [87].
- Future research could also look into integrating privacy-preserving techniques, like differential privacy, into the models to provide robustness against adversarial attacks while also ensuring the confidentiality of the data.
- As discussed, there are other strategies to generate adversarial perturbation. We aim to design a universal framework that is robust to the adversarial examples generated by any strategies.
- As quantum computing advances, it may offer new possibilities for developing more efficient and powerful adversarial defense mechanisms. Future research in this area could be highly fruitful.
- As we delve further into the realm of medical applications for artificial intelligence, it becomes increasingly evident that ensuring confidentiality and maintaining privacy will be imperative. As part of the future scope, there is a pressing need to develop and implement robust mechanisms and technologies that safeguard sensitive medical data. This includes the exploration of advanced encryption techniques, secure data-sharing protocols, and stringent access control measures. Additionally, the development of AI models that can operate effectively while preserving patient confidentiality and privacy will be a critical avenue of research. As we continue to harness the power of AI in healthcare, these endeavors will play a pivotal role in building trust and ensuring compliance with privacy regulations and standards.
- Lastly, testing the proposed models and defense mechanisms in real-world clinical settings would provide valuable insights into their practical applicability and performance, informing further refinement and development.

Author Contribution All authors contributed to the research and writing process of this paper. Each author has approved the submitted version of the paper.

Data Availability The datasets analyzed during the current study are available in the KAGGLE repository, https://github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md and https://www.kaggle.com/datasets/hgunraj/covidxct.

Code Availability The code generated during and/or analyzed during the current study is available from the corresponding author on reasonable request.

Declarations

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

Consent to Participate As this research did not involve human participants, consent to participate is not applicable.

Consent for Publication All authors agree to the terms of publishing and have given their consent to publish the content of this paper.

Competing Interest The authors declare that they have no competing interests.

References

1. West, C. P., Montori, V. M. & Sampathkumar, P. Covid-19 testing: The threat of false-negative results. Mayo Clin. Proceeding (2020).
2. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020:200432.
3. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, Yang C. Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. Journal of Medical Virology. 2020 .
4. Li D, Wang D, Dong J, Wang N, Huang H, Xu H, Xia C. False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: role of deep-learning-based CT diagnosis and insights from two cases. Korean J Radiol 2020;21(4):505–8.
5. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Xia L. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 2020:200642.
6. Sheridan C. Fast, portable tests come online to curb coronavirus pandemic. Nat Biotechnol. 2020.
7. Abdel-Zaher AM, Eldeib AM (2016) Breast cancer classification using deep belief networks. ExpertSyst Appl 46:139–144 .
8. Sun W, Tseng TB, Zhang J, Qian W (2017) Computerized medical imaging and graphics enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. ComputMed Imaging Graph 57:4–9.
9. D. C. Cires¸an, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in International Conference on Medical Image Computing and Computer-assisted Intervention, 2013, pp. 411–418.
10. Liu S, Liu S, Cai W,et al. Early diagnosis of Alzheimer's dis-ease with deep learning. In:International Symposium onBiomedical Imaging, Beijing, China 2014, 1015–18.
11. Brosch T, Tam R. Manifold learning of brain MRIs by deep learning.Med Image Comput Comput Assist Interv 2013;16:633–40
12. Majumdar A, Singhal V (2017) Noisy deep dictionary learning: application to Alzheimer's Disease classification. In: Neural networks (IJCNN), 2017 international joint conference on. IEEE, pp 2679–2683]
13. R. Li, W. Zhang, H. I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," Med Image Comput Comput Assist Interv, vol. 17, no. Pt 3, pp. 305–312, 2014.
14. Jin B, Che C, Liu Z, Zhang S, Yin X, Wei X (2018) Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access 6:9256–9261.
15. Meng, F., Kottlors, J., Shahzad, R. et al. AI support for accurate and fast radiological diagnosis of COVID-19: an international multicenter, multivendor CT study. Eur Radiol 33, 4280–4291 (2023). https://doi.org/10.1007/s00330-022-09335-9
16. Liang, H., Guo, Y., Chen, X. et al. Artificial intelligence for stepwise diagnosis and monitoring of COVID-19. Eur Radiol 32, 2235–2245 (2022). https://doi.org/https://doi.org/10.1007/s00330-021-08334-6
17. Alhasan, M., & Hasaneen, M. (2021, July). Digital imaging, technologies and artificial intelligence applications during COVID-19 pandemic. Computerized Medical Imaging and Graphics, 91, 101933. https://doi.org/10.1016/j.compmedimag.2021.101933
18. Hussain, M. A., Mirikharaji, Z., Momeny, M., Marhamati, M., Neshat, A. A., Garbi, R., & Hamarneh, G. (2022, December). Active deep learning from a noisy teacher for semi-supervised 3D image segmentation: Application to COVID-19 pneumonia infection in CT. Computerized Medical Imaging and Graphics, 102, 102127. https://doi.org/10.1016/j.compmedimag.2022.102127
19. Yin, M., Liang, X., Wang, Z. et al. Identification of Asymptomatic COVID-19 Patients on Chest CT Images Using Transformer-Based or Convolutional Neural Network–Based Deep Learning Models. J Digit Imaging (2023). https://doi.org/10.1007/s10278-022-00754-0.
20. Chen, H., Guo, S, Hao, Y. et al. Auxiliary Diagnosis for COVID-19 with Deep Transfer Learning. J Digit Imaging 34, 231–241 (2021). https://doi.org/10.1007/s10278-021-00431-8
21. Sheikh, B., Zafar, A. White-box inference attack: compromising the security of deep learning -based COVID-19 detection systems. Int J Inf Tecnol (2023). https://doi.org/10.1007/s41870-023-01538-7
22. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Jamalipour Soufi, G. (2020, October). Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Medical Image Analysis, 65, 101794. https://doi.org/10.1016/j.media.2020.101794
23. Venkataramana, L., Prasad, D.V.V., Saraswathi, S. et al. Classification of COVID-19 from tuberculosis and pneumonia using deep learning techniques. Med Biol Eng Comput 60, 2681–2691 (2022). https://doi.org/10.1007/s11517-022-02632-x
24. Singh, M., Bansal, S., Ahuja, S. et al. Transfer learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data. Med Biol Eng Comput 59, 825–839 (2021). https://doi.org/10.1007/s11517-020-02299-2
25. Mamalakis, M., Swift, A. J., Vorselaars, B., Ray, S., Weeks, S., Ding, W., Clayton, R. H., Mackenzie, L. S., & Banerjee, A. (2021, December). DenResCov-19: A deep transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from X-rays. Computerized Medical Imaging and Graphics, 94, 102008. https://doi.org/10.1016/j.compmedimag.2021.102008
26. Huang, M. L., & Liao, Y. C. (2022, November). Stacking Ensemble and ECA-EfficientNetV2 Convolutional Neural Networks on Classification of Multiple Chest Diseases Including COVID-19. Academic Radiology. https://doi.org/10.1016/j.acra.2022.11.027.
27. Menon, S., Mangalagiri, J., Galita, J. et al. CCS-GAN: COVID-19 CT Scan Generation and Classification with Very Few Positive Training Images. J Digit Imaging (2023). https://doi.org/10.1007/s10278-023-00811-2
28. Yuan, J., Wu, F., Li, Y. et al. DPDH-CapNet: A Novel Lightweight Capsule Network with Non-routing for COVID-19 Diagnosis Using X-ray Images. J Digit Imaging (2023). https://doi.org/10.1007/s10278-023-00791-3
29. Di, D., Shi, F., Yan, F., Xia, L., Mo, Z., Ding, Z., Shan, F., Song, B., Li, S., Wei, Y., Shao, Y., Han, M., Gao, Y., Sui, H., Gao, Y., & Shen, D. (2021, February). Hypergraph learning for identification of COVID-19 with CT imaging. Medical Image Analysis, 68, 101910. https://doi.org/10.1016/j.media.2020.101910
30. Kiziloluk, S., Sert, E. COVID-CCD-Net: COVID-19 and colon cancer diagnosis system with optimized CNN hyperparameters using gradient-based optimizer. Med Biol Eng Comput 60, 1595–1612 (2022). https://doi.org/10.1007/s11517-022-02553-9
31. Wang, T., Nie, Z., Wang, R. et al. PneuNet: deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis

using Vision Transformer. Med Biol Eng Comput 61, 1395–1408 (2023). https://doi.org/10.1007/s11517-022-02746-2
32. Shang, Y., Wei, Z., Hui, H. et al. Two-stage hybrid network for segmentation of COVID-19 pneumonia lesions in CT images: a multicenter study. Med Biol Eng Comput 60, 2721–2736 (2022). https://doi.org/10.1007/s11517-022-02619-8
33. Chamberlin, J. H., Aquino, G., Schoepf, U. J., Nance, S., Godoy, F., Carson, L., Giovagnoli, V. M., Gill, C. E., McGill, L. J., O'Doherty, J., Emrich, T., Burt, J. R., Baruah, D., Varga-Szemes, A., & Kabakus, I. M. (2022, August). An Interpretable Chest CT Deep Learning Algorithm for Quantification of COVID-19 Lung Disease and Prediction of Inpatient Morbidity and Mortality. Academic Radiology, 29(8), 1178–1188. https://doi.org/10.1016/j.acra.2022.03.023.
34. Fang, C., Bai, S., Chen, Q., Zhou, Y., Xia, L., Qin, L., Gong, S., Xie, X., Zhou, C., Tu, D., Zhang, C., Liu, X., Chen, W., Bai, X., & Torr, P. H. (2021, August). Deep learning for predicting COVID-19 malignant progression. Medical Image Analysis, 72, 102096. https://doi.org/10.1016/j.media.2021.102096
35. Xue, W., Cao, C., Liu, J., Duan, Y., Cao, H., Wang, J., Tao, X., Chen, Z., Wu, M., Zhang, J., Sun, H., Jin, Y., Yang, X., Huang, R., Xiang, F., Song, Y., You, M., Zhang, W., Jiang, L., . . . Xie, M. (2021, April). Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information. Medical Image Analysis, 69, 101975. https://doi.org/10.1016/j.media.2021.101975
36. sheikh, B., Zafar, A. Beyond accuracy and precision: A robust deep learning framework to enhance the resilience of face mask detection models against adversarial attacks. Evol Syst (2023). https://doi.org/10.1007/s12530-023-09522-z
37. Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., Moon, S., Lim, J. K., & Ye, J. C. (2022). Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. Medical Image Analysis, 75, 102299. https://doi.org/10.1016/j.media.2021.102299
38. Sheikh, B., Zafar, A. RRFMDS: Rapid Real-Time Face Mask Detection System for effective COVID-19 monitoring. SN COMPUT SCI 4, 288 (2023). https://doi.org/10.1007/s42979-023-01738-9
39. Bertolini, M., Brambilla, A., Dallasta, S. et al. High-quality chest CT segmentation to assess the impact of COVID-19 disease. Int J CARS 16, 1737–1747 (2021). https://doi.org/10.1007/s11548-021-02466-2
40. Fang, X., Kruger, U., Homayounieh, F. et al. Association of AI quantified COVID-19 chest CT and patient outcome. Int J CARS 16, 435–445 (2021). https://doi.org/10.1007/s11548-020-02299-5
41. Wang, R., Jiao, Z., Yang, L. et al. Artificial intelligence for prediction of COVID-19 progression using CT imaging and clinical data. Eur Radiol 32, 205–212 (2022). https://doi.org/10.1007/s00330-021-08049-8.
42. Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T. N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., Halm, N., El Hajj, S., Bompard, F., Neveu, S., Hani, C., Saab, I., Campredon, A., Koulakian, H., Bennani, S., Freche, G., Paragios, N. (2021, January). AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. Medical Image Analysis, 67, 101860. https://doi.org/10.1016/j.media.2020.101860 .
43. Dong, J., Wu, H., Zhou, D. et al. Application of Big Data and Artificial Intelligence in COVID-19 Prevention, Diagnosis, Treatment and Management Decisions in China. J Med Syst 45, 84 (2021). https://doi.org/10.1007/s10916-021-01757-0.
44. Santosh, K.C. COVID-19 Prediction Models and Unexploited Data. J Med Syst 44, 170 (2020). https://doi.org/10.1007/s10916-020-01645-z.
45. Bae, H., Jang, J., Jung, D., Jang, H., Ha, H., Lee, H., & Yoon, S. (2018). Security and privacy issues in deep learning. arXiv:1807.11655.
46. Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv:1712.05526.
47. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning-based medical image analysis systems. Pattern Recognition, 110, 107332.
48. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).
49. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
50. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
51. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv:1412.6572.
52. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083.
53. [Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In Artificial intelligence safety and security (pp. 99–112). Chapman and Hall/CRC.
54. Ahmed, N., Natarajan, T., & Rao, K. (1974, January). Discrete Cosine Transform. IEEE Transactions on Computers, C–23(1), 90–93. https://doi.org/10.1109/t-c.1974.223784.
55. Akter, S., Shamrat, F. M. J. M., Chakraborty, S., Karim, A., & Azam, S. (2021, November 13). COVID-19 Detection Using Deep Learning Algorithm on Chest X-ray Images. Biology, 10(11), 1174. https://doi.org/10.3390/biology10111174
56. Wu, X., Chen, C., Zhong, M., Wang, J., & Shi, J. (2021, February). COVID-AL: The diagnosis of COVID-19 with deep active learning. Medical Image Analysis, 68, 101913. https://doi.org/10.1016/j.media.2020.101913
57. Meng, Y., Bridge, J., Addison, C., Wang, M., Merritt, C., Franks, S., Mackey, M., Messenger, S., Sun, R., Fitzmaurice, T., McCann, C., Li, Q., Zhao, Y., & Zheng, Y. (2023, February). Bilateral adaptive graph convolutional network on CT based Covid-19 diagnosis with uncertainty-aware consensus-assisted multiple instance learning. Medical Image Analysis, 84, 102722. https://doi.org/10.1016/j.media.2022.102722
58. Gao, K., Su, J., Jiang, Z., Zeng, L. L., Feng, Z., Shen, H., Rong, P., Xu, X., Qin, J., Yang, Y., Wang, W., & Hu, D. (2021, January). Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Medical Image Analysis, 67, 101836. https://doi.org/10.1016/j.media.2020.101836
59. Li, G., Togo, R., Ogawa, T. et al. COVID-19 detection based on self-supervised transfer learning using chest X-ray images. Int J CARS 18, 715–722 (2023). https://doi.org/10.1007/s11548-022-02813-x
60. Qi, X., Brown, L.G., Foran, D.J. et al. Chest X-ray image phase features for improved diagnosis of COVID-19 using convolutional neural network. Int J CARS 16, 197–206 (2021). https://doi.org/10.1007/s11548-020-02305-w
61. gifani, P., Shalbaf, A. & Vafaeezadeh, M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. Int J CARS 16, 115–123 (2021). https://doi.org/10.1007/s11548-020-02286-w
62. Brinati, D., Campagner, A., Ferrari, D. et al. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning:

A Feasibility Study. J Med Syst 44, 135 (2020). https://doi.org/10.1007/s10916-020-01597-4.
63. Wang, S., Kang, B., Ma, J. et al. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). Eur Radiol 31, 6096–6104 (2021). https://doi.org/10.1007/s00330-021-07715-1.
64. Xu, B., Martín, D., Khishe, M. et al. COVID-19 diagnosis using chest CT scans and deep convolutional neural networks evolved by IP-based sine-cosine algorithm. Med Biol Eng Comput 60, 2931–2949 (2022). https://doi.org/10.1007/s11517-022-02637-6.
65. Younis, M. C. (2021, June). Evaluation of deep learning approaches for identification of different corona-virus species and time series prediction. Computerized Medical Imaging and Graphics, 90, 101921. https://doi.org/10.1016/j.compmedimag.2021.101921.
66. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. arXiv 2013, arXiv:1312.6199
67. H. Hirano, K. Koga, K. Takemoto, Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks, PLoS One 15 (12) (2020.
68. Wang, L., Lin, Z. Q., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports, 10(1), 1-12.
69. Rahman, A., Hossain, M. S., Alrajeh, N. A., & Alsolami, F. (2020). Adversarial examples—Security threats to COVID-19 deep learning systems in medical IoT devices. IEEE Internet of Things Journal, 8(12), 9603-9610.
70. J. Fei, Z. Xia, P. Yu, and F. Xiao, "Adversarial attacks on fingerprint liveness detection," EURASIP J. Image Video Process., vol. 2020, no. 1 pp. 1–11, 2020, doi: https://doi.org/10.1186/s13640-020-0490-z.
71. Kakizaki, K., & Yoshida, K. (2019). Adversarial image translation: Unrestricted adversarial examples in face recognition systems. arXiv:1905.03421.
72. Pal, B., Gupta, D., Rashed-Al-Mahfuz, M., Alyami, S. A., & Moni, M. A. (2021). Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for covid-19 prediction from chest radiography images. Applied Sciences, 11(9), 4233.
73. Qi, G., Gong, L., Song, Y., Ma, K., & Zheng, Y. (2021). Stabilized medical image attacks. arXiv:2103.05232.
74. Gongye, C., Li, H., Zhang, X., Sabbagh, M., Yuan, G., Lin, X. & Fei, Y. (2020). New passive and active attacks on deep neural networks in medical applications. In Proceedings of the 39th international conference on computer-aided design (pp. 1–9).
75. Gougeh, R. A. (2021). How Adversarial attacks affect Deep Neural Networks Detecting COVID-19?.
76. Sheikh, B.U.h., Zafar, A. Untargeted white-box adversarial attack to break into deep leaning based COVID-19 monitoring face mask detection system. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-15405-x
77. L. (2022, February 17). GitHub - lindawangg/COVID-Net: COVID-Net Open Source Initiative. GitHub. https://github.com/lindawangg/COVID-Net.
78. T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," Computers in Biology and Medicine, p. 103792, 2020.
79. A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images," Computer Methods and Programs in Biomedicine, p. 105581, 2020.
80. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," Physical and Engineering Sciences in Medicine, p. 1, 2020.
81. L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images," arXiv, pp. arXiv–2003, 2020.
82. Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp) (pp. 39–57). Ieee.
83. M. H. Wani and A. R. Faridi, Deep Learning-Based Video Action Recognition: A Review. 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2022, pp. 243–249. https://doi.org/10.1109/ICCCIS56430.2022.10037736.
84. K. Roshan, Zafar, A, sheikh, B.U.H. Untargeted White-box Adversarial Attack with Heuristic Defence Methods in Real-time Deep Learning based Network Intrusion Detection System, Computer Communications, 2023, ISSN 0140-3664. https://doi.org/10.1016/j.comcom.2023.09.030
85. S. B. Ul Haque, A. Zafar and K. Roshan, "Security Vulnerability in Face Mask Monitoring System," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 231-237
86. K. Roshan, A. Zafar and S. B. Ul Haque, "A Novel Deep Learning based Model to Defend Network Intrusion Detection System against Adversarial Attacks," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 386-391
87. sheikh, B.U.H., Zafar, A. Unlocking adversarial transferability: A security threat towards deep learning-based surveillance systems via black box inference attack- a case study on face mask surveillance. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-16439-x.