IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 February 2024, accepted 21 February 2024, date of publication 26 February 2024, date of current version 1 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369900

RESEARCH ARTICLE

Fast and Efficient Lung Abnormality Identification With Explainable AI: A Comprehensive Framework for Chest CT Scan and X-Ray Images

MD. ZAHID HASAN[1], SIDRATUL MONTAHA[2], INAM ULLAH KHAN[1], MD. MEHEDI HASSAN[3], (Member, IEEE), ABDULLAH AL MAHMUD[1], A. K. M. RAKIBUL HAQUE RAFID[1], SAMI AZAM[4], (Member, IEEE), ASIF KARIM[4], (Member, IEEE), SPYRIDON PROUNTZOS[5], EFTHYMIA ALEXOPOULOU[5], UMAMA BINTA ASHRAF[6], AND SHEIKH MOHAMMED SHARIFUL ISLAM[7]

[1]Health Informatics Research Laboratory (HIRL), Department of Computer Science and Engineering, Daffodil International University, Dhaka 1341, Bangladesh
[2]Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada
[3]Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh
[4]Faculty of Science and Technology, Charles Darwin University, Casuarina, NT 0909, Australia
[5]Second Department of Radiology, Faculty of Medicine, Attikon University General Hospital, National and Kapodistrian University of Athens, 115 27 Athens, Greece
[6]Department of Radiology and Imaging, BIRDEM General Hospital, Dhaka 1000, Bangladesh
[7]Institute for Physical Activity and Nutrition, Deakin University, Melbourne, VIC 3125, Australia

Corresponding authors: Asif Karim (asif.karim@cdu.edu.au) and Sheikh Mohammed Shariful Islam (shariful.islam@deakin.edu.au)

ABSTRACT A novel automated multi-classification approach is proposed for the anticipation of lung abnormalities using chest X-ray and CT images. The study leverages a publicly accessible dataset with an insufficient and unbalanced number of images, addressing this issue by employing the data augmentation approach DCGAN to balance the dataset. Various preprocessing procedures are applied to improve features and reduce noise in lung pictures. As the base for the model, the vision trans-former and convolution-based compact convolutional transformer (CCT) model is utilized. To determine the best model configuration, an ablation study is performed on the original CCT model using a CT scan dataset with image dimensions of 32 × 32. Following that, this model is trained on the X-ray dataset to evaluate performance on an entirely other modality. The performances are compared to six pre-trained models with 32 × 32 images. While traditional models achieved modest performance, with test accuracies ranging from 43% to 77% and 49% to 73% requiring lengthy training times, the suggested model performed exceptionally well, obtaining test accuracies of 99.77% and 95.37% for CT and X-ray, respectively with a short training duration of 10-12 and 40-42 seconds/epoch. Robustness is demonstrated through the progressive reduction of the number of training images, with findings indicating that the model maintains good performance even on a reduced dataset. An explainable AI technique Grad-CAM is used to explain the model's judgment. Grad-CAM-based color visualization is shown to explain model assessments and help health specialists make quick, confident decisions. This study used image preprocessing and deep learning techniques to detect lung anomalies, and it addressed the challenges of training time and computational complexity.

INDEX TERMS Lung disease, chest x-ray, CT scan, image preprocessing, compact convolutional transformer, deep convolutional GAN, explainable AI.

I. INTRODUCTION

Respiratory disorders account for 5 of the top 30 causes of mortality, and early prevention, disease control, and effective

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

treatment are crucial [1]. Accurate diagnosis is critical for recovery and optimal chances of survival. Chest radiography and computed tomography (CT) scans are two techniques of imaging that are widely used to identify lung ailments. Chest X-rays are less costly, easier to use, more commonly available, and faster than CT scans. They contain a great deal of information on the patient's condition [2]. However, interpreting is a challenge, especially for non-radiology trained doctors. An experienced radiologist may even face difficulties in distinguishing lung pathology. On the other hand, CT scans give extremely fine resolution in three dimensions, but they require significantly higher radiation levels and more expensive equipment than ordinary X-ray imaging [3].

Numerous researches have demonstrated that using deep learning algorithms trained on routine chest radiography pictures can detect and classify lung diseases with high accuracy. Convolutional Neural Networks (CNNs), in particular, have produced outstanding results in the identification of many illnesses, including lung disorders [4]. These models, however, rely on the accessibility of massive volumes of labeled training data or require fine-tuning from pre-trained CNNs of millions of parameters [5]. Vision Transformer (ViT) [6], a self-attention [7] based model inspired by natural language processing (NLP) was first introduced in computer vision tasks. The backbone used was a pure transformer architecture. These models produce moderate accuracies, slightly lower than ResNets of similar size when trained on mid-sized datasets like ImageNet without heavy regularization. This disappointing result occurs because transformers lose several of the inductive biases of CNNs, including translation equivariance and localization. As a consequence, they do not simplify fit when trained on inadequate data, though the situation varies when the models are trained on more datasets (14M-300M images). According to the findings of [6], inductive bias is overcome by large-scale training. ViT achieves excellent results when pre-trained appropriately. However, ViT is a data-hungry approach that has rendered transformers unusable for a wide range of important tasks as many areas are scarce. It also requires a lot of computing power. Hassani et al. [8] introduce a Compact Convolutional Transformer (CCT) model that enables sequential pooling and replaces patch embedding with convolutional embedding, allowing for a greater inductive bias and generating positional embedding to avoid the big data requirements of Compact Transformers. This improves efficiency and makes input image sizes more flexible, while demonstrating less reliance on Positional Embedding than the other models.

Labeled data are frequently scarce in the medical imaging field, particularly for advanced diseases [5]. Due to these limitations, supervised learning techniques may struggle to perform well with new data. In contrast, unsupervised visual learning techniques rely solely on unlabeled data. Autoencoders and Generative Adversarial Networks (GAN) are examples of unsupervised learning methods that can be used to augment data [9]. Autoencoders receive data as input and return a compressed representation of that data, but the reconstructed image is frequently blurry and of poor quality [2]. GANs, on the other hand, can autonomously identify patterns in the input data to generate new instances derived from the original dataset. Variational autoencoders (VAEs) are a specific form of autoencoder that can be used as an example of a generative model. They are especially useful for generating new samples that are similar to the original data and can help increase the size and diversity of the training set, thereby enhancing the efficacy of machine learning models. Explainable AI (XAI) is the development of AI models and algorithms to make its judgments and predictions more clear and comprehensible. Grad-Cam, which visualizes the classification process to increase model transparency, is a common way to show these capabilities [10].

In this study, we investigate the identification and categorization of lung disorders from pre-processed CT scans and chest X-rays utilizing the same CCT model. The study considered two diverse modality datasets e.g-COVID-19 Radiography and CT scan. The dataset of COVID-19 Radiography comprises four different classes of COVID-19, Opacity of Lung, and Pneumonia (Normal and Viral). With the exception of lung opacity, the same classes are present in the CT-scan dataset. As images of different modalities have different imaging protocols and dissimilar characteristics in the affected region's shape and sizes, artifacts and noises, classifying diseases with high accuracy using the same framework for both datasets can be challenging. The study attempts to evolve a robust framework that is applies to both types of datasets.

The main contributions of this study are:

- Addressing class imbalance in medical imaging datasets using a GAN, which enhances the performance and reliability of classification models.
- Determining diverse image preprocessing techniques through extensive experimentation, that maintain high image quality while preserving critical diagnostic information.
- Proposing CTXNET, a model optimized for chest X-ray and CT scan datasets, potentially improving diagnostic accuracy and efficiency of lung abnormalities. After that, Grad-Cam is used to visualize the model's classification.
- Achieving reduced training time by integrating convolutional blocks into the vision transformer, enabling efficient processing of low-resolution images thus contributing to both space and time complexity.
- Providing a comprehensive comparison of the proposed CTXNET model with six transfer learning networks for evaluating the effectiveness of the model particularly in terms of precision and training time.

Accurate diagnosis of lung disorders such as COVID-19 and other lung infections such as pneumonia (viral/bacterial) has been given much attention in recent years. These CAD approaches have been developed to aid clinicians in analyzing

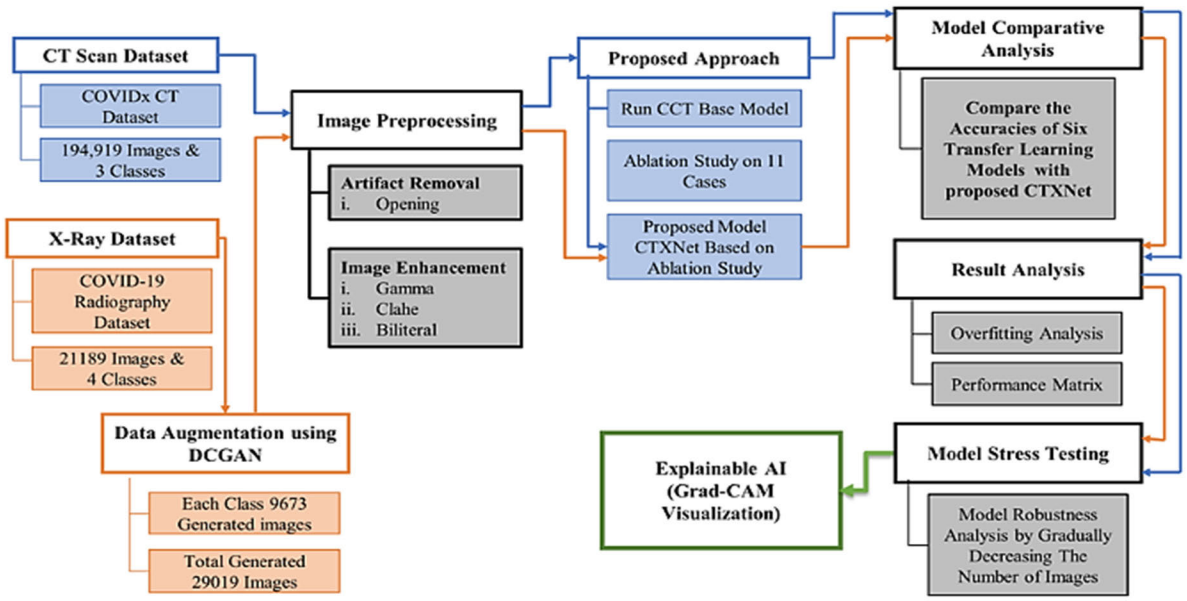


FIGURE 1. The entire process of developing CTXNET architecture for classifying Lung CT scans and X-rays.

medical images. Deep learning algorithms are increasingly popular due to their advantage in automated feature mining and their high recognition accuracy. Abiyev and Maaitah [11] recommended a deep-learning-based method to identify lung diseases using 112,120 frontal-view X-ray images. They produced a computerized CNN model for detecting chest diseases and demonstrated that the CNN model outperformed alternative soft computing approaches regarding training accuracy, testing accuracy, and training duration. They applied their model to 12 classes and achieved 92.4% test accuracy using CNN model for this multi-classification of diseases. Varela-Santos and Melin [12], used computer vision and soft computing approaches to analyze pneumonia from chest X-rays. The Region of Interest (ROI) was retrieved employing X-ray image segmentation, observed by feature extraction, also a neural network was applied for classifying the data, achieving 90% accuracy on the ChestXRay14 database. Wang et al. [13] suggested a new respiratory disease detection technique based on computer vision algorithms and a collaborative CNN model. The segmentation method was used to detect the ROI in lung images for successful pneumonia categorization, and global and regional characteristics were retrieved. The categorization was done with the help of a cooperative CNN model and 92.3% accuracy was achieved on the RSNA Pneumonia Detection Competition dataset. Thakur et al. [14] demonstrated a deep learning based strategy employing a CNN variation named VGG16 to classify pneumonia using chest X-ray images. They employed the transfer-learning and modified technique and attained 90.54% accuracy. InstaCovNet-19, an integrated stacking model, was developed by Gupta et al. [15]. They employed a number of transfer learning models which were used to create a stack-like design. The suggested model obtains a classification accuracy of 99.08% on three classes (Pneumonia, COVID-19, Normal) and an accuracy of 99.53% on two classes (COVID, NON-COVID). Jain et al. [16] used an imbalanced database with 1345 regular patients, 3632 pneumonia instances, also, 490 COVID-19 instances divided into 3 groups. They investigated different architectures and discovered that the Xception architecture had the top accuracy of 97.97%. A fine-tuned AlexNet model was created using a Support Vector Machine (SVM) framework to classify lung disorders [17]. Ouchicha et al. [18] proposed CVDNet, a residual neural network, in order to differentiate pneumonia and normal classes from COVID-19. They also made use of a fivefold cross-validation method to assess their system and reached an average accuracy of 96.69%. Ozturk et al. [19] suggested DarkCovidNet, a system that can detect lung diseases automatically that works on binary class classification with 98.08% accuracy and solves multi-class classification as well. Nevertheless, pre-processing procedures were not applied to the X-ray images in their research.

Using axial lung CT-scan images, Modegh et al. [20] suggested a novel interpretable deep neural network to categorize healthy persons, COVID-19 patients, and various pneumonia disorders patients. In addition, the algorithm recognizes diseased locations and determines the proportion of infected lung volume. There was a dataset of 3359 samples taken away from six different medical institutes to test and instruct the model. In differentiating healthy from unhealthy people and COVID-19 from other diseases, the network attained sensitivity of 97.75% and 98.15%, and specificity of 87% and 81.03%, respectively. Li et al. [21] suggested an au-to-encoder-based architecture for distinguishing positive from negative covid-19 instances and achieved an accuracy

of 94.7%; however, they only employed a small number of images to train the model. Xu et al. [22] employed ResNet18 with location attention to identify COVID-19 patients by CT-scan images. Their entire accuracy level was just 86.7%, which is not optimal. COVID-19 has been identified in certain trials using a combination of lung CT scans and X-rays. In this paper [23], the authors proposed a modified deep neural network with a comprehensive accuracy of 96.13% for chest X-rays and 95.83% for Computed tomography by well-adjusted datasets. A standard modified VGG-19 approach was presented in this study [24] to recognize and characterize COVID-19 with the help of X-ray images and CT scans. In another study, researcher [25] developed a pre-learned approach for differentiating COVID-19 instances from non-COVID-19 instances, with 82.94% accuracy for the dataset of CT-scan and 93.94% for the dataset of chest x-rays. In this research [26], the authors utilized an imbalanced dataset of X-ray and CT-scan pictures in terms of identifying COVID-19 from streptococcus and SARS virus contagions. They used a modified VGG19 model, an InceptionV3 architecture, and a decision tree classifier through the modified VGG19 model. The validation accuracy of the model was 91%. For identification of the COVID-19 patients from chest X-ray and CT scan pictures, Sedik et al. [27] used a combination of machine learning and deep learning methods. The authors used two data-augmentation strategies to gain the effectiveness of deep learning methods depending on Convolutional Long Short-Term Memory and CNN. There are drawbacks to utilizing machine learning (ML) systems, for instance complication, overfitting, and low accuracy when training with imbalanced datasets. However, most of the studies described above did not work with both CT and X-ray datasets. Limitations include in carrying out suitable image preprocessing techniques, data augmentation and model hyperparameter tuning. In this research, these impediments are given attention as developing a single framework with high interpretation capabilities for two different modalities is quite a challenging task.

II. PROPOSED METHODOLOGY

As the major goal of this research is developing a single CAD framework for two different modalities, two datasets have been employed for the experiments. Figure 1 depicts the workflow for classifying lung diseases, which includes image preprocessing and selecting appropriate models.

The X-Ray dataset has been augmented as it has few images compared to the CT scan image dataset. Data Augmentation is a strategy for reducing overfitting in deep learning models, resulting in a more extensive range of available data. The Deep Convolutional Generative Adversarial Network (DCGAN) method is applied here to augment the X-ray images for individual classes. To get the best performance from deep learning models, artifacts are removed from both datasets using similar image preprocessing approaches while simultaneously boosting image contrast and quality.

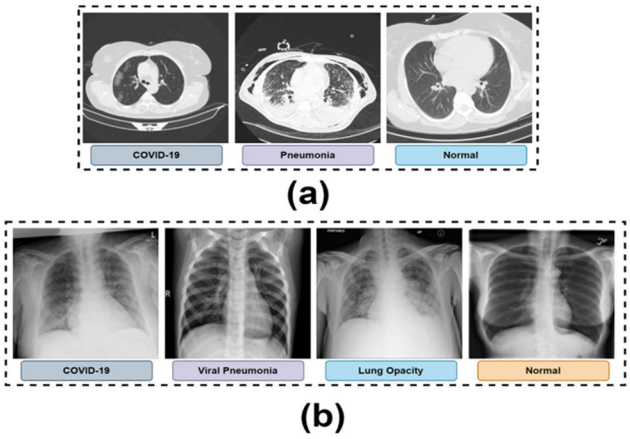


FIGURE 2. Images from each class of the (a) CT Scan Image (b) Chest X-ray image.

In the image preprocessing step, morphological opening is applied first to remove the artifacts from the images of both datasets. Afterwards, gamma correction, Contrast Limited Adaptive Histogram Equalization (CLAHE) and bilateral filter are used to upgrade image contrast as well as quality. Several statistical techniques, named Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Root Mean Squared Error (RMSE) are applied to ensure that the images are preprocessed efficiently without the quality being affected. Afterwards, the CT scan dataset images are fed to the original CCT model, which is considered the root model. The proposed CTXNET model is constructed by an ablation study on the Base CCT model. Before applying the model to the dataset of X-rays, the optimal configuration had been acquired. The accuracy of the proposed prototype has been contrasted with six different transfer learning models. Overfitting and performance matrix analysis are done for both datasets. In addition, the model robustness is tested by gradually decreasing the quantity of images from the test dataset. To represent the importance or relevance of different parts of the input image to the model's decision, Grad-CAM based visualization is shown at the end.

III. DATASET DESCRIPTION

We have worked with two different datasets: Chest X-ray and CT scan. The CXR dataset is accessed from COVID-19 Radiography Dataset from ''Kaggle'' comprising a total image of 21149 images [28]. There are 4 classes: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. The COVID-19 class includes 3616 images, Lung Opacity contains 6012 images, Normal contains 10192 images, and 1345 images are found for Viral Pneumonia (Figure 2-b). This dataset is imbalanced in terms of image numbers in the different classes. A description of both datasets is given in Table 1.

The CT scan dataset is a 2D form of 3D images. Therefore, it contains several 2D slices of each patient. The dataset is collected from [29] which comprises 194919 images in

TABLE 1. Description of the used datasets.

Name	X-ray Image Dataset	Computerized Tomography Dataset
Amount of Images	21149	194919
Dimension	299×299	512×512
Covid-19	3616	94545
Normal	10192	60083
Lung Opacity	6012	-
Viral Pneumonia	1345	40291

three classes: COVID-19, Normal and Viral Pneumonia. The COVID-19 class includes 94545 images, Normal has 60083 images, and Viral Pneumonia has 40291 images (Figure 2-a). We have used single slice CT images of each patient for testing the data.

The COVID-19 virus is commonly observed on CT scans, as consolidations (high-attenuation regions) and/or ground-glass opacities (hazy regions) with a peripheral distribution pattern. In X-ray, initially there may be few or no abnormalities observed. It may eventually manifest as infiltrates or irregular opacities in the lungs.

In viral pneumonia, a CT scan may reveal consolidations, patchy ground-glass opacities, and a peribronchial distribution pattern. An X-ray examination may reveal infiltrates or patchy opacities in the lungs, similar to COVID-19.

Lung opacities on an X-ray image may manifest as elevated density regions, which indicate atypical lung tissue or fluid. The cause and characteristics of the opacities need to be further evaluated [30].

Although, the clinical symptoms of COVID-19 and influenza pneumonia are very similar. The precautions required of the general population and health workers to prevent transmission, disease management methods, and prognosis, on the other hand, are considerably different. As a result, distinguishing COVID-19 from influenza pneumonia in the early stages of infection is critical for a well-timed and suitable therapeutic plan. There are imaging differences between influenza and COVID-19 that allow the two to be distinguished [31]. Here, Figure 3 depicts the indication of the infected area on X-ray images of normal, COVID-19, lung opacity, and non-COVID viral pneumonia. According to the source of the data, they are annotated by experts and in our paper a medical doctor has confirmed the annotations of each lung class.

IV. DATA AUGMENTATION

The chest X-ray dataset has a data imbalance issue and insufficient images to train a vision transformer-based model. Therefore, to increase the volume of the dataset and address the data imbalance problem, a data augmentation technique called GAN is employed. While traditional augmentation techniques generate data based on geometrical or photometrical approaches, GAN does not change the geometry or intensity level of a picture.

In the analysis of medical images, a common problem is a limited number of images in the datasets. Deep convolutional networks have been demonstrated to be useful in the process of medical image including detection, classification and segmentation when dealing with sufficient amounts of data. Due to rare illnesses, the privacy of patients, and the demand of medical professionals for labeling, including the cost; in addition, laboring effort is required to undertake medical imaging operations, underscoring the difficulties in creating large medical image datasets [29]. GAN [9], a data augmentation technique is gaining popularity among deep learning researchers, particularly in the computer vision sector [9]. Substantial progress has been achieved in overcoming difficulties in terms of realistic image creation, data scarcity and image-to-image translation [32]. Despite this, using GANs for real world issues continues to encounter major hurdles, such as (1) the creation of good-quality images, (2) image variety, and (3) consistent training. GAN's shortcomings include learning supervision, the inaptitude to identify overfitting, and instability when applied to small data sets [33]. We therefore employ DCGAN, which chains GAN and CNN, providing a robust architecture through alteration [34].

A. DCGAN

DCGAN's design and operation are identical to the original GAN. Using a min-max algorithm, GAN is a proficient deep learning based generative network that creates synthetic pictures with higher diversity without guidance. The system's training process is further enhanced by artificial data generated using a generative model which adds variety and enriches the dataset. As a result, CNN models have increased generalization capabilities, reducing overfitting problems [35]. DCGAN, an improved augmentation strategy that addresses the constraints of traditional data augmentation approaches, may deceive the generator into learning the distributions of the augmented data, which may diverge from the distribution of the source data [36]. It combines two neural networks, the generator and the discriminator; these two neural networks cooperate to produce new data samples by reducing the gap between the probability distributions of the source and the produced data. The generator figures out how to create new information with similar characteristics as the preparation set, while the discriminator will try to recognize genuine and produced test data in a given set. GANs can interpret source-manufactured images and create similar images for training. By lowering the functional loss due to training, the network generator enhances its capacity to produce synthetic data [37], [38]. On the other hand, the discriminator improves its ability to distinguish between authentic and synthetic data by maximizing a related loss

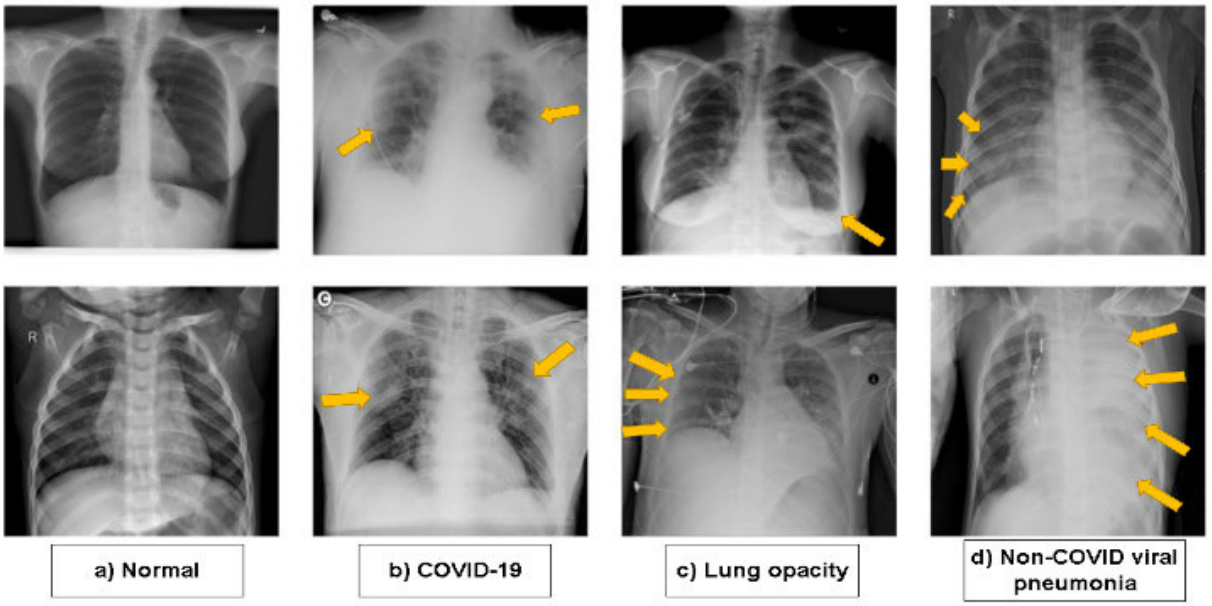


FIGURE 3. Image data examples for multiclass classification to detect (a) normal, (b) COVID-19, (c) lung opacity, and (d) non-COVID viral pneumonia. The yellow markers denote the location that has been infected.

function. To train the generator and discriminator networks, the equation is-

$$\min_{N}\max_{M} V_{GAN}(M,N)=\mathbb{E}_{x\sim Pdata(x)}\left[\log M(x)\right] + \mathbb{E}_{z\sim Pz(z)}\left[\log\left(1-M\left(N\left(z\right)\right)\right)\right] \quad (1)$$

where M refers to the Discriminator and N is the Generator, and V_{GAN} is the anticipated values of large genuine and fake occurrences, x signifies one-of-a-kind data, and is the probability that x came from the special data conveyance instead of from the generated data dissemination is the arbitrary clamor variable inspected from a standard typical distribution, the total plan of the generator arrange is delineated in Figure 4.

All images have been downsized to 224×224 before using DCGAN. The generator starts with a random input of 100×1 noise vector, which is then added to the dense layer and transformed to $14 \times 14 \times 512$. In this architecture, we employ 4 conv2D transpositions and 1 conv2D layer to up sample an image size from $14 \times 14 \times 512$ to $224 \times 224 \times 1$.

Data of size $14 \times 14 \times 512$ is transformed into an image of size $28 \times 28 \times 256$ after passing through the first Convolutional2D Transpose. The architecture is the same in the following three layers; the output from the first Conv2D transpose layer passes via the batch normalization layer and Conv2D transpose and becomes re-shaped in each layer respectively to $56 \times 56 \times 128$, $112 \times 112 \times 64$, and $224 \times 224 \times 32$. Using the conv2D layer in the last layer, we receive an output with an image size of $224 \times 224 \times 1$.

The simulated data produced by the generator feeds forward via the network in the discriminator system. The source images of the dataset and the produced images from the generator network are sent to the discriminator. Next, a combination of four-block convolution layers is applied where each block includes a dropout layer, a LeakyReLu activation function, and Conv2D. The discriminator functions act as a binary classifier using binary cross-entropy, predicting the probability of the images being fake or real.

The discriminator will mistakenly identify the fake picture as real if the created image is very similar to an actual one. But if the generator creates a duplicate picture which is not similar to the actual images, the discriminator recognizes it as fake and through backpropagation, the generator's weights are changed. The generator now has more accurate weights, and it keeps attempting to deceive the discriminator by identifying fake images as authentic. A robust generator competent in making untrue pictures that closely take after genuine pictures can be utilized to extend the number of pictures in a certain dataset through this generator and discriminator cycle [39], [40].

The discriminator in a DCGAN differentiates between genuine and fake images by learning to extract features from input images that are representative of their content. Then, these features are input into a classification layer, which generates a probability score indicating whether the image is authentic or fabricated. The discriminator is trained with a binary cross-entropy loss function that penalizes it for incorrectly classifying genuine images as fake or vice versa. As the generator develops over time, distinguishing between authentic and fabricated images becomes more challenging for the discriminator. DCGANs have been shown to generate high-quality medical images, and a radiologist verified the generated images. The generated images are closely similar but not identical, and original data were limited, which is why

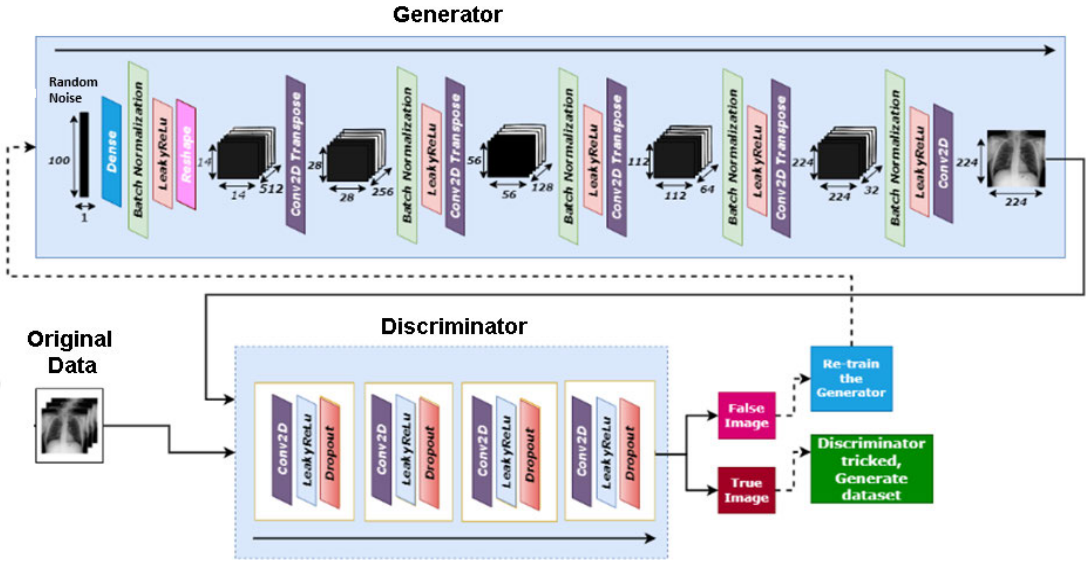


FIGURE 4. Architecture of DCGAN.

they also have been used for train, validation and testing in this study.

B. TRAINING SCHEME OF GAN

The Normal class has the most significant number of images (10192). We have balanced the number of images near the Normal class of 10192 for the other three classes. At first, the images are resized to 224 × 224 which are further used for training the DCGAN model. A learning rate of 0.0008, batch size of 128, and the optimizer Adam are applied to train the model. The epoch number is set according to the image number in the initial dataset. It is often observed that a deep learning model tends to perform well while employing a totally adjusted dataset. The robustness of the model can be validated fairly using a slightly imbalanced dataset. Therefore, while augmenting the dataset, the number of images in the classes is kept slightly different. In this regard, for COVID-19, after augmentation, the images are raised from 3616 to 13289, for Lung Opacity, 6012 to 13911 and for Viral Pneumonia 1345 to 12128. This way, the total number of images is increased from 21165 to 49520 after augmentation.

Figure 5 illustrates the primary images and images generated by DCGAN resulting in almost similar images for both generated and root images.

V. IMAGE PREPROCESSING

Before providing images to neural networks, image preprocessing is a crucial step. An adequate classification performance for both modalities may be impossible to achieve since understanding the features of images is difficult owing to the complicated structure and presence of distortions and noises [41]. Several image preprocessing methods have been used on both datasets to draw out object noises and adjust the image quality. The parameter values of the algorithms are selected based on the optimal output after extensive testing with the images. All preprocessing procedures, including artifact removal and enhancement of images, have been described in sequence in this section.

Figure 6 illustrates the entire image preprocessing techniques with the outputs for both datasets.

A. MORPHOLOGICAL OPENING

Morphological opening [42] is employed to remove artifacts from the images. Morphological opening is the process that completely eliminates single-pixel artifacts like noisy spikes and tiny spurs and blackens small objects. Objects frequently maintain their original dimensions and forms. In this operation, the first procedure is erosion, followed by dilation on the input image. A 5∗5 kernel size removes texts and artifacts in the background region of the CT and X-ray images while the object remains unchanged.

B. GAMMA CORRECTION

Gamma correction can be used to modify an image's overall brightness. It can be used on photographs that have been excessively dark or bleached out [43]. Pixel intensity values should be expanded and compressed for darker and fading photos. A gamma value of 1.2 is chosen after testing with different gamma values, as illustrated in Figure 6.

C. CLAHE

The images have gray levels distributed via the histogram equalization (HE) approach. Hence, the likelihood of each gray level is identical. To increase image quality, HE fine-tunes the intensity levels of dark and low-contrast images. Adaptive Histogram Equalization is an improved form of HE. Thus, instead of utilizing the image's global data, it upgrades local contrast and edges in each region. AHE, in spite of the fact that, more-over upgrade the noise components of

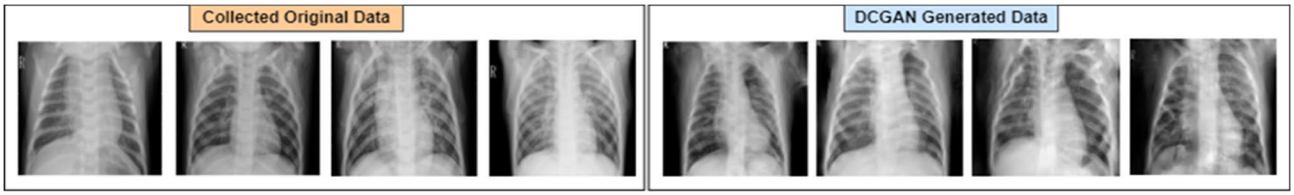


FIGURE 5. Original image and GAN-Produced image.

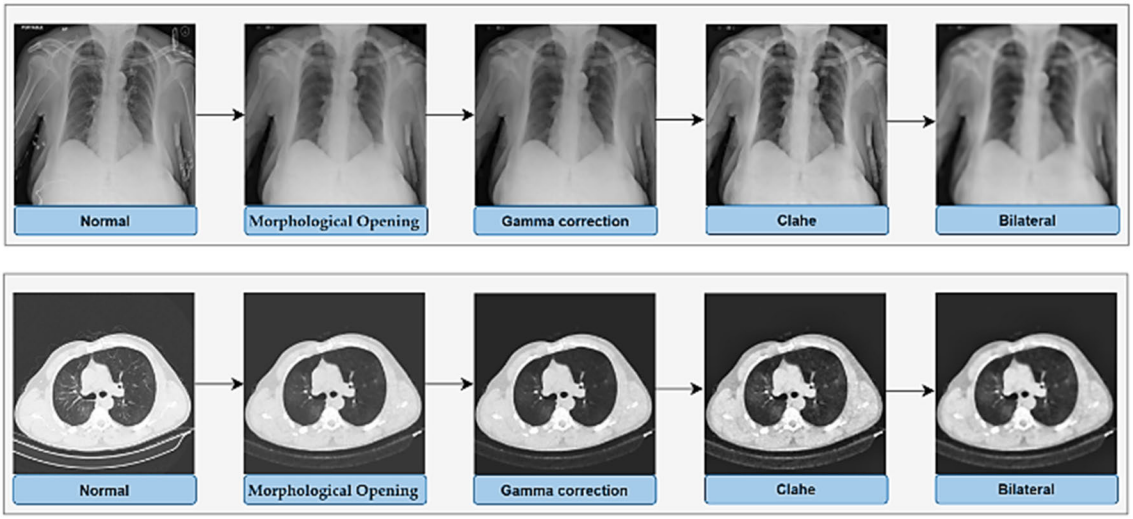


FIGURE 6. Original image and GAN-Produced image.

images progressed with CLAHE. The HE approaches can oversaturate some areas when used on medical images. The identical procedure as AHE is utilized by CLAHE to address this problem.

D. BILATERAL FILTER

A bilateral filter is a robust filtering implementation that makes full use of the spatial statistics among pixel locations and the local pixel value. This filtering method merges two components, Euclidean distance and radiometric separation, to resolve the problem of Gaussian obscurity in traditional image filtering approaches. The method of applying the algorithm is depicted in Figure 6, where the output images are found smooth and have the edge preserved.

E. VERIFICATION

In certain circumstances, the image quality may be degraded as a result of using the techniques described above. We determine the MSE, SSIM, PSNR, and RMSE values by contrasting the original and pre-processed images to verify that image quality did not fall. MSE is the average squared, whereas RMSE is the average difference between expected and actual values. PSNR estimates picture or video quality by comparing it to a reference image or video, whereas SSIM evaluates image similarity by comparing structural information such as brightness, contrast, and structure. MSE and RMSE have been evaluated to obtain a more comprehensive picture of the algorithm's efficacy. In general, the MSE range is 0 to 1, with a value of more than 0.5 indicating a high-quality image. The SSIM value varies between -1 and 1. A score of close to 1 implies that quality is intact, and a value near to 1 suggests that the processing step has little effect on image quality [44]. For 8-bit images, the ideal PSNR varies from 30 to 50 decibels (dB). A value less than 20dB is considered inadequate [45]. The RMSE estimates the difference between the original and processed images, and a lower RMSE value indicates higher image quality. We applied MSE, PSNR, SSIM, and RMSE on 49,520 X-ray images, although displaying all these data in the table is inconvenient. We therefore chose ten photos at random to represent the values. However, the average MSE for chest X-ray images is 0.37, PSNR is 32.4, SSIM is 0.993, and RMSE is 0.54. Figure 7 also includes a pie chart that depicts the percentage of photos based on the range of PSNR values for all of 49,520 images. For calculating MSE, PSNR, SSIM, and RMSE of CT scan images, 25% of images are considered as this dataset contains ample an amount of data. The average MSE for CT scan images is 32.00, PSNR is 32.2, SSIM is 0.98, and RMSE is 0.61. Table 2 and Table 3 display the values for twenty randomly selected pictures from the X-ray and the CT scan datasets respectively.

Figure 7(a) shows that nearly 28% of images are in the range between 31.01 and 32.00, which is a large proportion of the total X-ray images, 26% of images are between 32.01 and 33.00, 20% of images are in the range of 33.01-34.00, 9.32% images are between 38.01 and 39.00, and 24% images are

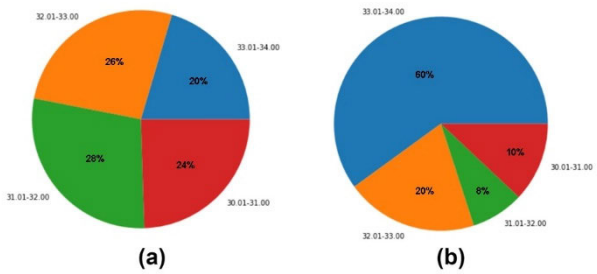


FIGURE 7. Pie chart of PSNR value for (a) X-ray images and (b) CT scan images.

TABLE 2. Statistical values for ten Chest X-ray images.

Image	PSNR	SSIM	MSE	RMSE
1	32.81	0.9946	0.34	0.58
2	33.01	0.9949	0.32	0.56
3	32.48	0.9917	0.36	0.60
4	32.43	0.9950	0.37	0.60
5	32.22	0.9950	0.38	0.62
6	32.87	0.9940	0.33	0.57
7	33.10	0.9947	0.31	0.56
8	31.97	0.9960	0.41	0.64
9	32.03	0.9975	0.40	0.63
10	32.61	0.9949	0.35	0.59

TABLE 3. Statistical values for ten CT Scan Images.

Image	PSNR	SSIM	MSE	RMSE
1	31.68	0.9930	0.44	0.66
2	32.19	0.9954	0.39	0.62
3	31.98	0.9936	0.41	0.64
4	32.86	0.9916	0.33	0.57
5	31.56	0.9924	0.45	0.67
6	32.84	0.9955	0.33	0.57
7	32.24	0.9941	0.38	0.61
8	33.22	0.9950	0.30	0.54
9	32.09	0.9944	0.40	0.63
10	32.01	0.9939	0.40	0.63

between 30.01 and 31.00. Figure 7(b) About 60% of the images are in the range between 33.01 and 34.00, which is a significant proportion of the CT scan images, 20% of images are between 32.01 and 33.00, 8% are between 31.01 and 32.00 and 10% are between 30.01 and 31.00.

As the goal is to develop a model that can accurately identify lung abnormalities, pre-processing methods including contrast enhancement, and noise reduction can improve the model's performance by enhancing the features relevant to detecting lung abnormalities. If we use the original/actual image to evaluate the statistical values, they will achieve a total score as reference images will be real ones. Tables 2 and 3 show that the attained statistical values are higher than the standard values, indicating that the quality of the images

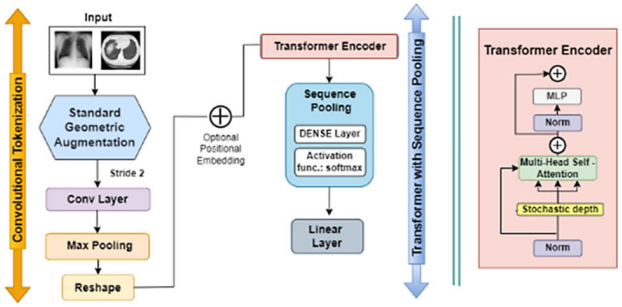


FIGURE 8. Structure of CCT.

after the processing is preserved. So, it is possible to conclude that the majority of images fall within allowable thresholds. Even after employing several pre-processing techniques, the image quality is effectively retained.

VI. PROPOSED MODEL

By outperforming traditional CNN models in terms of computing efficiency and training time, ViT has become well-known in the computer vision domain. ViTs' encoder-decoder blocks enable the processing of numerous sequential data sets simultaneously in significantly less time. They can use their self-attention mechanism to identify long-distance links between successive items. Because of this, they do exceptionally well in photo categorization tasks [46], [47], [48]. Most medical datasets are insufficient for adequate operation of ViTs since ViTs need large amounts of data for training. To address this problem, CCT, a fusion of ViT with convolution, is presented [8]. With a local receptive field that keeps up the local data of the image, CCT utilizes CNN blocks as patching blocks. The self-attention strategy catches associations concerning parts of the image and combines all related data.

A. COMPACT CONVOLUTIONAL TRANSFORMER

A transformer with sequential pooling and Convolutional Tokenization are the two major building components of CCT systems. The mechanism of CCT is illustrated in Figure 8.

Patches of the input images are produced using the Convolutional Tokenization block. The patches of these images are combined into a sequence. Tokenization is the process of splitting an image into smaller pieces called tokens. The convolutional tokenization processes for a given image x is:

$$x_0 = \text{MaxPool}(\text{ReLU}(\text{Conv2D}(x))) \tag{2}$$

where, only 2 of the 64 filters in the convolutional layer (Conv2d) are equipped with the ReLU. The Conv2D feature maps are then scaled down by the max pool layer. The input picture size for the convolutional tokenization block may be variable.

The output patches produced by the first block are then sent to the transformer-based backbone, where the encoder block is made up of a Multihead Self-Attention (MSA) layer and a

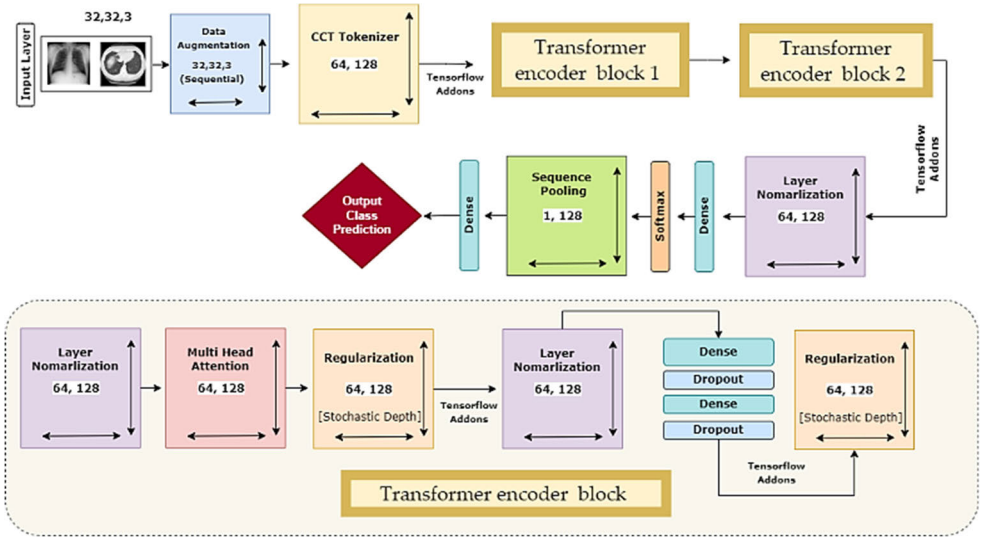


FIGURE 9. Base CCT model architecture.

Multilayer Perceptron (MLP) head. The transformer encoder employs dropout, GELU activation, and layer normalization (LN). Positional Embed-ding and Self-attention (including Multi-head Self-attention) enable the model to recognize image spatial relationships. Self-attention lets the model process images by focusing on different sections. Multi-head Self-attention helps the model focus on distinct subsets of features inside each patch to capture more complex feature interactions. The sequence pooling layer pools the output of the transformer backbone, using sequence pooling [11]. With the help of this sequence pooling, the network may evaluate the sequential embedding's of latent space produced by the encoder and enhance data correspondence for the input. The sequence pooling layer pools the full sequence of data since it contains pertinent data from diverse input image regions. This process is known as mapping transformation.

After the sequence pooling part, the images are then classified by after passing through the linear layer.

B. BASE CCT ARCHITECTURE

This work proposes a model named CTXNET by modifying the original CCT, which is accomplished by conducting ablation study on the original CCT model. The Base CCT architecture is shown in Figure 9.

An input layer, data augmentation layer with various geometric augmentation methods, CCT tokenizer, multi-head attention layers, regularization layers, pooling layers, dropout layers, dense layers, and output dense layers are all components of the CCT architecture. The input images have the dimensions of $32 \times 32 \times 3$ where the data augmentation layer performs. The CCT Tokenizer block receives the enhanced pictures as input, and the resulting image is then resized to 64×128. In the tokenizer block of the convolutional layer, the size of stride and kernel are set as 2 and 4 respectively, along with the kernel size of 4 for the pooling layer. Tokenization is followed by tensorflow additions before the data is sent to the transformer encoder block. The first layer normalization, multi-head attention, regularization and second layer normalization are followed by two sets of dense and dropout layers having a dropout ratio of 0.1, forming the layers in a specified order. The transformer encoder block's last layer is linked to another regularization layer. This regularization layer is used to regularize the output, which has a size of 64×128. Another transformer-encoded block similar to the previous one is then applied. Two layers - one for regularization and one for normalization - are then applied to. The dense layer with softmax function creates outputs having dimension of 64×1, the normalized output is created. This is passed on to a layer called sequence pooling, which produces output data with a dimension of 1×128. The chest X-ray pictures are finally divided into four groups using a linear classification layer.

The model is trained for 100 epochs with a learning rate of 0.001, batch size of 128 and optimizer Adam. Categorical Cross entropy is chosen as the loss function.

C. ABLATION STUDY

As previously mentioned, to optimize the performance, we conducted ablation research on this CCT network by modifying the layer design and tuning the values of hyper parameters. In this regard, eleven ablation studies are carried out. After all ablation studies are finished, the suggested CTXNET network is developed having a more robust design, better performance, and faster processing time.

D. CTXNET NETWORK

An illustration of the architecture of CTXNET network is shown in Figure 10.

The proposed CTXNET design is made more efficient to reduce the duration of training, maximize performance, and less time complexity. The final CTXNET design, which includes fewer transformer encoder blocks than the base

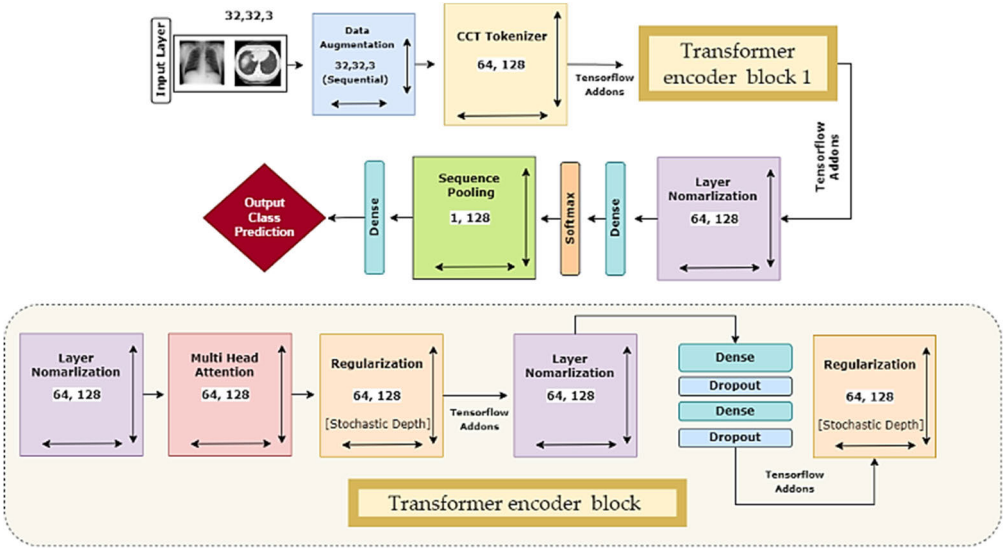


FIGURE 10. Proposed model CTXNET architecture.

CCT model (see section VII-B). As shown in Figure 10, the CTXNET model has a single encoder block, whereas the original CCT architecture has two. This makes the model shallow and allows for quicker training. With the exception of a few modifications to the model's hyper parameters, such as stride and kernel, the other components of the architecture remain the same.

Contrary to transformer-based models, this model does not require positional encoding, which aids in keeping the computing cost low. The computational complexity of self-attention is O (n^2.d), where n is the span of the input sequence and d is the number of dimensions. The computational complexity increases with the addition of positional encoding (O (n^2.d+d.n^2)). The training phase of the model is shorter and needs fewer resources since positional encoding is not required in the CTXNET model and the transformer only depends on the self-attention mechanism. As a result, the model is much more efficient.

E. TRANSFER LEARNING MODELS

To assess the performance of our proposed approach more rigorously, a comparative analysis with six TL models named VGG16, VGG19, ResNet50, ResNet50V2, ResNet152 and MobileNet which are trained with the same datasets and the performance, including training duration, is recorded. To train the models, the strategy is kept the same as the CCT model.

1) VGG ARCHITECTURE

The VGG model comprises three fully linked layers after the first five blocks of convolutional layers. The VGG16 contains 16 weighted layers and VGG19 contains 19 weighted layers and both have the drawback of being expensive to assess and require a large memory resource. VGG16 contains around 138 million parameters while VGG19 contains about 143 million parameters.

2) RESNET ARCHITECTURE

Deep convolutional networks called Residual Networks (ResNets) [49] are based on the principle of skip-ping convolutional blocks utilizing shortcut relations to create blocks known as residual blocks.

Essential components of the ResNet-50 architecture adhere to two design principles: every layer filter number is constant for a similar feature map and doubles if the output feature map size is halved. The network concludes with fully connected layers activated using softmax. There are 50 weighted layers in all, with 23,534,592 learnable parameters.

ResNet152's main innovation was its ability to train highly complicated neural network models with more than 150 layers. ResNet is regarded as the best deep learning architecture since superior results can be tweaked and generated easily because it has many networks and layers of architecture which have a considerable timing complexity.

ResNet50V2 is the improved version of the ResNet50 [49]. The propagation of the links between blocks in ResNet50V2 has been altered.

3) MOBILENET ARCHITECTURE

According to Sandler et al. [50], MobileNet V2 enhances the performance of mobileNet models within the different model ranges, workloads, and criteria. MobileNet's concept is to swap convolutional layers that can be separated depending on the depth. It produces almost the same results as ordinary convolution but is much faster. Standard 3 × 3 convolution is the first step in the MobileNetV1 architecture, followed by 13 depth-wise separable convolutional blocks.

F. GRAD CAM BASED VISUALIZATION

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique used in explainable AI (XAI) to display and comprehend deep neural network decisions. It creates

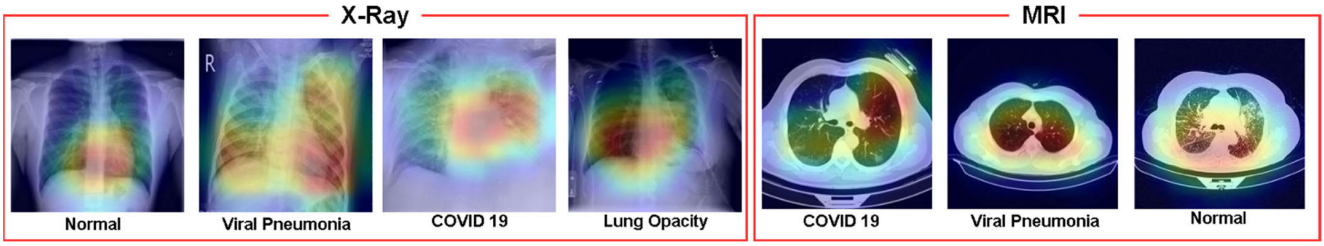


FIGURE 11. Grad-CAM based heatmap visualization.

a heatmap that emphasizes the relevant regions in an input image that helped the model forecast. It entails using the model's output as input and tracing the gradient down to the last convolutional layers. These layers consist of the most detailed information discovered by the model prior to classification. Grad-CAM's heatmap colors show the importance or relevance of various elements of the input image to the model's judgment. Warmer colors (for example, red) signify greater importance, while cooler colors (for example, blue) suggest less importance [51]. We can learn more about which portions of the image influenced the model's output and better understand its decision-making process by visualizing these heatmaps.

Three procedures are used to generate the model's heat map: gradient calculation, averaging gradients to compute alphas, and the final Grad-CAM heat map calculation. To begin, the gradient of a given output neuron (y^c) with respect to the activation (A^k) of the convolutional layers is computed. Interestingly, because the input image also represents the feature map, the value of a specific gradient in Equation (4) is equal to the input image.

$$Computed\ Gradient = \frac{\delta y^c}{\delta y^c} \tag{3}$$

The next step involves determining the alpha value by taking the average of a group of global variables with respect to the breadth measure "I" and the height dimension index "j". This computation yields the neuron importance weights, represented as α_k^c in Equation (4).

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\delta y^c}{\delta y^c} \tag{4}$$

Feature map activation A^k is executed as a weighted combination in the final phase. From the alpha values computed in the preceding phase, these weights are derived. The ultimate Grad-CAM thermal map is produced by the computation that ensues. This heat map is subjected to a ReLU operation, which, as shown in Equation (6), retains only the positive values and sets all negative values to 0.

$$L^C_{Grad-Cam} = ReLU(\sum_k \alpha_k^c A_k) \tag{5}$$

The resultant output is a normalized abrasive heatmap intended for visual representation [52]. It is illustrated visually in Figure 11.

VII. RESULT AND DISCUSSION

A. EVALUATION METRICS

Several measures, including accuracy (ACC), recall, precision, specificity and F1-score are calculated to evaluate the efficiency of the proposed classification model. For additional statistical examination of the model, the false positive rate (FPR), the false negative rate (FNR), the false discovery rate (FDR), the negative predicted value (NPV) and Matthews Correlation Coefficient (MCC) are also assessed. These evaluation metrics are produced using true negative (TN), true positive (TP), false negative (FN) and false positive (FP) values that are obtained from the confusion metrics [53].

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{RecalI} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{F}_1 = 2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{11}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{12}$$

$$\text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{14}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

B. EVALUATION METRICS

All ablation studies carried out for this research are described in this section. By changing individual elements of the base architecture, various experiments are carried out and the performance of the altered model is examined. A total of eight ablation studies are conducted for this study. The findings of these ablation investigations are listed in Tables 4 and 5 showing test accuracy and training time.

TABLE 4. Ablation studies regarding transformer layers, loss functions, kernel size and stride size.

Ablation Study 1: changing transformer layers						
Configuration	transformer layers	Parameter no.	Epoch x training time	Total training time	accuracy	Finding
1	3	0.57M	100 x 122s	200-210 minutes	93.73%	Highest training period
2	2	0.47M	100 x 85s	140-145 minutes	93.45%	Large training period
3	1	0.24M	100 x 42s	60-70 minutes	93.35%	Lowest training period

Ablation Study 2: Altering Loss Function					
Configuration	Loss Function	Parameter no.	Epoch x training time	accuracy	Finding
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy
3	Mean Squared Error	0.3M	100 x 42s	89.15%	Accuracy dropped
4	Mean absolute error	0.3M	100 x 42s	88.55%	Accuracy dropped

Ablation Study 3: changing kernel size					
Configuration	Loss Function	Parameter no.	Epoch x training time	accuracy	Finding
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy
3	Mean Squared Error	0.3M	100 x 42s	89.15%	Accuracy dropped
4	Mean absolute error	0.3M	100 x 42s	88.55%	Accuracy dropped

TABLE 4. *(Continued.)* Ablation studies regarding transformer layers, loss functions, kernel size and stride size.

Ablation Study 4: Changing Stride Size					
Configuration	Loss Function	Parameter no.	Epoch x training time	accuracy	Finding
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy
3	Mean Squared Error	0.3M	100 x 42s	89.15%	Accuracy dropped
4	Mean absolute error	0.3M	100 x 42s	88.55%	Accuracy dropped

- ***Study 1: Changing transformer layers***

In order to attain the best accuracy, the transformer layer number contained in the primary model is changed by adjusting or removing encoded blocks. The outcomes are various numbers of encoded blocks are shown in Table 4. The model is able to attain an accuracy of 93.35% with a noticeably shorter training period while using configuration 3. The training times for the remaining two configurations are 140-145 and 200-210 minutes, respectively, whereas configuration 3 only required 60–70 minutes. As configuration 3 has the least trainable parameters and minimum training time per epoch this is picked for further experiments.

- ***Study 2: Altering the loss function***

To achieve optimal performance, experiments are done with four individual loss functions. The Categorical Cross entropy loss function exhibits the greatest test accuracy of 93.35% (Table 4). Therefore, Categorical Cross entropy is chosen.

- ***Study 3: Altering kernel size***

Several transformer layer kernel sizes are investigated in search of the model's most ideal setup wherewith a kernel size 3 resulting in the maximum test accuracy, of 93.99% (Table 4). The architecture requires the least training time per epoch, only 42 seconds.

Study 4: Changing stride size

Table 4 displays the results for stride sizes of 1, 2, 3, and 4. With a stride size of 1, the model's accuracy improved to 97.9% while sustaining a per-epoch training period of 42 seconds.

- ***Study 5: Changing the type of pooling layer***

Two different types of pooling layers (max and average) are used in the experiments (Table 5). The maximum test

TABLE 5. Ablation studies regarding the pooling layer, activation function, optimizer and learning rate.

Case Study 5: Changing the type of pooling layer					
Configuration	**Loss Function**	**Parameter no.**	**Epoch x training time**	**accuracy**	**Finding**
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy

Case Study 6: Changing activation function					
Configuration	**Loss Function**	**Parameter no.**	**Epoch x training time**	**accuracy**	**Finding**
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy
3	relu	0.24M	100 x 42s	99.63%	Maximum accuracy
4	softmax	0.24M	100 x 42s	48.51%	Accuracy dropped
5	prelu	0.24M	100 x 42s	97.9%	Previous accuracy

Case Study 7: Changing Optimizer					
Configuration	**Loss Function**	**Parameter no.**	**Epoch x training time**	**accuracy**	**Finding**
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy
3	relu	0.24M	100 x 42s	99.63%	Maximum accuracy
4	softmax	0.24M	100 x 42s	48.51%	Accuracy dropped
5	prelu	0.24M	100 x 42s	97.9%	Previous accuracy

TABLE 5. *(Continued.)* Ablation studies regarding the pooling layer, activation function, optimizer and learning rate.

Case Study 8: Learning rate					
Configuration	**Loss Function**	**Parameter no.**	**Epoch x training time**	**accuracy**	**Finding**
1	Binary Crossentropy	0.3M	100 x 42s	91.36%	Accuracy dropped
2	Categorical Crossentropy	0.3M	100 x 42s	93.35%	Maximum accuracy
3	relu	0.24M	100 x 42s	99.63%	Maximum accuracy
4	softmax	0.24M	100 x 42s	48.51%	Accuracy dropped

accuracy of 97.9% is obtained with the max pooling layer, so this is selected for additional investigations.

- ***Study 6: Changing activation function***

A classification model's effectiveness is influenced by the activation function. The performance of a network can be enhanced by choosing the best activation function. Six activation functions were investigated as shown in Table 5. ReLU performs the best, with 99.63% test accuracy and a per epoch time of 42 seconds. Therefore, the activation function is adopted.

- ***Study 7: Changing the optimizer***

Experiments are conducted with five different optimizers. The optimizers' learning rates were set at 0.001. Table 5 shows that the Adam optimizer resulted in the greatest test accuracy of 99.63%.

- ***Study 8: Changing the learning rate***

Table 5 shows the outcomes of testing of Adam optimizer with diverse learning rates: 0.01, 0.006, 0.001, and 0.0008. The best performance is achieved with a learning rate of 0.001, yielding a 99.78% test accuracy while keeping training time/epoch at 42 seconds.

Figure 12 illustrates how test accuracy gradually increased throughout the ablation studies carried out on the basic model.

C. PERFORMANCE ANALYSIS OF CTXNET MODEL

The proposed CTXNET model is produced by completing ablation experiments on the base model, which provides improvement for the classification accuracy. This is accomplished by modifying and configuring the model in different ways using the CT Scan dataset. This proposed system is also trained and tested on X-ray dataset to evaluate its

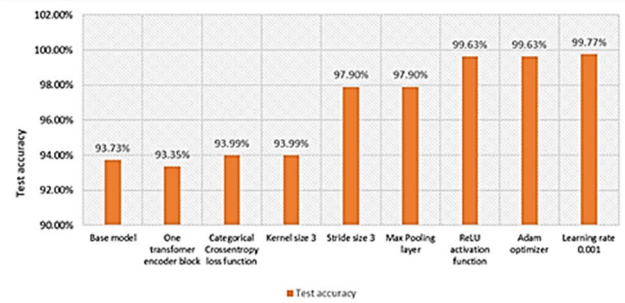


FIGURE 12. Gradual increase of test accuracy over eight ablation studies.

TABLE 6. Performance evaluation matrices of the CTXNET model for the CT scan and the Chest X-ray dataset.

Measure	CT scan dataset	X-Ray dataset
Recall	99.68%	94.87%
Specificity	99.87%	98.31%
Precision	99.79%	94.79%
NPV	99.86%	98.29%
FPR	0.012%	0.169%
FDR	0.025%	0.052%
FNR	0.031%	0.051%
F1 Score (F1)	99.71%	94.83%
MCC	0.995%	0.931%

performance on a dataset which it is not optimized for. While training on chest X-ray dataset, the configuration was kept the same. Table 6 displays evaluation metrics for the proposed CTXNET model, for both of the datasets. Table 6 represents the results of CT scan and X-ray dataset in terms of several performance metrics.

When testing with the CT scan dataset, the suggested CTXNET model obtains an F1 score of 99.71%, recall and specificity scores of 99.68% and 99.87%, respectively, and a precision of 99.79%. FPR and FNR values were 0.012% and 0.031%, respectively which is remarkably low. The model has an FDR of 0.025% with an NPV of 99.86%. The model's MCC value is 0.99%.

The model also gives promising results when trained and tested with the X-ray dataset yielding an F1 score of 94.83% and a specificity of 98.31%. FPR and FNR values are 0.169% and 0.051% respectively and the NPV value is 98.29%. In terms of FDR and MCC, the model performs quite well with a low FDR value of 0.052% and an MCC value of 93%. Across all evaluation measures, quite similar outcomes between the two datasets can be observed. Despite being optimized only for the CT scan dataset, the model is able to yield a comparable classification performance on the X-ray dataset, demonstrating the robustness of the model.

Figure 13 and 14 showcase the accuracy and loss curves of the CTXNET model, when trained on the CT scan and X-ray datasets respectively.

When testing with the CT scan dataset, the suggested CTXNET model obtains an F1 score of 99.71%, recall and

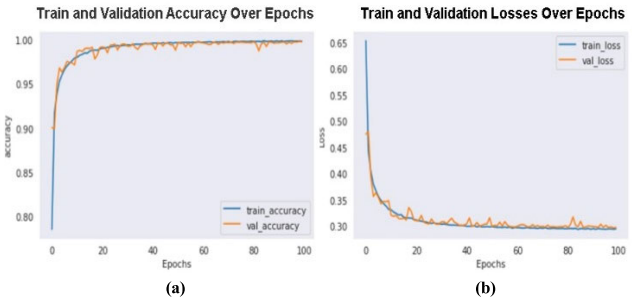


FIGURE 13. Loss curve and accuracy curve of CTXNET model while trained on CT scan dataset.

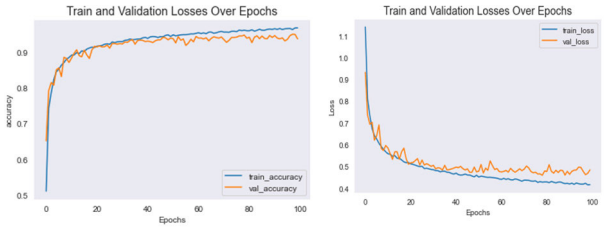


FIGURE 14. Loss curve and accuracy curve of CTXNET model while trained on X-ray dataset.

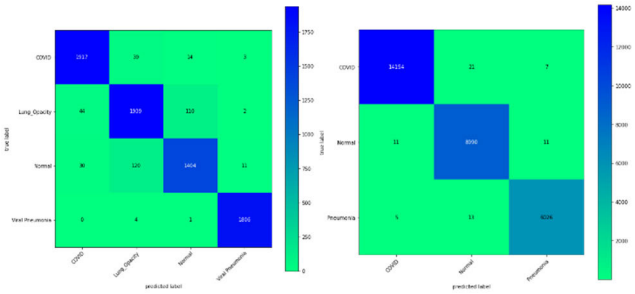


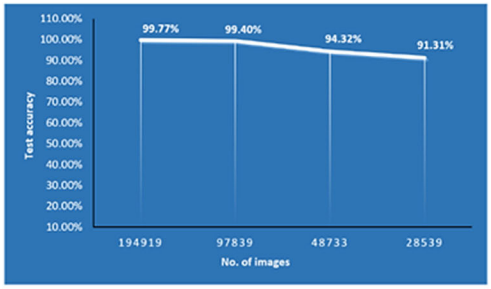
FIGURE 15. Confusion matrix of the proposed CTXNET model for (a) the X-ray and (b) the dataset CT scan dataset.

specificity scores of 99.68% and 99.87%, respectively, and a precision of 99.79%. FPR and FNR values were 0.012% and 0.031%, respectively which is remarkably low. The model has an FDR of 0.025% with an NPV of 99.86%. The model's MCC value is 0.99%.

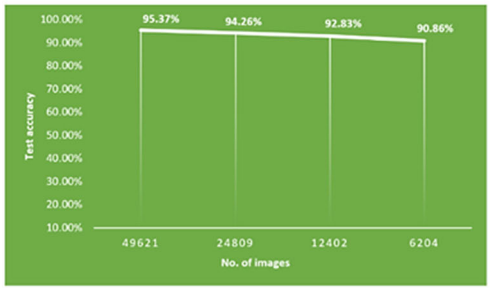
There is no evidence of overfitting during the model's training process as no significant gap is found in the training and validation curves for both datasets (Figure 13 and 14). Correspondingly, the loss curves for both datasets exhibit the same pattern. It can be said that neither overfitting nor under fitting occurred during the model's training phase on either dataset.

The confusion matrices for both datasets produced by the CTXNET model are displayed in Figure 15.

The test images' true labels are indicated by row values where column values serve as representations for the labels the model predicted on the test set. The test image numbers that the model correctly predicted are listed as diagonal values within the confusion matrices (Figure 15). However, the model is not biased toward any particular class. The model makes almost comparable numbers of accurate predictions for each class, further demonstrating the model's robustness.



(a)



(b)

FIGURE 16. Evaluation of the proposed model for (a) the Chest CT scan dataset and (b) the Chest X-ray dataset with a reduced number of images.

In addition, experiments are carried out by gradually reducing the volume of the dataset in order to evaluate the consistency of the proposed model in terms of classification performance. The CTXNET model is trained and tested in several stages. In each stage, the total image number in the dataset was decreased to roughly half of what it was previously. This experiment is carried out from both datasets. The results are visualized in Figure 16.

For the CT scan dataset, Figure 16(a) indicates that even after training the model with half (97839) of the image number as the primary dataset, the accuracy decreases by less than 1%. Reducing the image number further to 48733 results in a moderate test accuracy of 94.32%. Utilizing 28539 images to train and test the model still yields a reasonable performance with a test accuracy of 91.31%.

With similar experiments conducted on the X-ray dataset (Figure 16-b), the model is able to achieve 93.26% test accuracy with half the samples (24809 images), which means that a drop in accuracy of approximately 1% was observed. Moreover, with 12402 images, a moderate test accuracy of 91.08% is obtained. When using only 6204, the test accuracy decreased to 87.86%. For conventional CNN and ViT models, 6,000 images are considered quite a small number for training. However, even with a minimal number of images (6,000), the proposed CTXNET model generates a good outcome while also having short training times.

D. PERFORMANCE COMPARISON WITH TRANSFER LEARNING(TL) MODELS

The input picture dimensions are kept at 32 × 32 pixels for all six models and they are trained and assessed with both of the datasets. In this regard, every model undergoes 100 training epochs. Table 7 showcases the performance of the models for both datasets.

TABLE 7. Performance comparison of CT scan and X-ray dataset with six transfer learning CNN models utilizing an image size of 32 × 32 pixels.

Model	Number of params	CT scan Dataset			X-ray dataset		
		Per epoch time (seconds)	Total time (hours)	Test accuracy	Per epoch time (seconds)	Total time (hours)	Test accuracy
VGG19	2026436	170-173	05:20 - 05:50	73.84 %	61-63	01:50 - 02:00	76.97 %
VGG16	14716740	170-172	05:20 - 05:50	72.26 %	61-63	01:50 - 02:00	79.51 %
ResNet152	58379140	170-174	05:20 - 05:50	58.82 %	61-63	01:50 - 02:00	53.39 %
ResNet50	23595908	170-173	05:20 - 05:50	60.53 %	61-63	01:50 - 02:00	67.77 %
ResNet50V2	23572996	170-175	05:20 - 05:50	63.22 %	61-63	01:50 - 02:00	65.35 %
MobileNet	3232964	170-175	05:20 - 05:50	49.34 %	61-63	01:50 - 02:00	43.42 %
CTXNET	**241861**	**40-42**	**01:05 - 01:10**	**99.77 %**	**11-12s**	**0:18-0:20**	**95.3%**

The findings of Table 7 show that, VGG16 obtained the maximum accuracy of 79.51% on the X-ray dataset, whereas VGG19 achieved the highest accuracy, of 73.84%, on the CT scan dataset. The accuracy of the other models ranged from 43% to 77% for both datasets. It is evident that CNN models struggle to perform even moderately when trained on input images with small dimensions such as 32 × 32 pixels. On the other hand, CTXNET stands out as being particularly resilient, surpassing all six cutting-edge CNN models with peak test accuracies of 99.77% and 95.37% respectively for the CT scan dataset and the X-ray dataset with input images of 32 × 32 pixels. Compared to cutting-edge CNN models, the CTXNET model's 241,861 trainable parameters are low. Because there are fewer parameters, training takes less time (10)–12 seconds per epoch for both datasets), compared to CNN models which have more parameters (over 60 and 170 seconds per epoch for the X-ray and CT scan datasets respectively). This reduces the overall training time for the X-ray dataset (49621 images) from approximately two hours to only 18 to 20 minutes and for the CT scan dataset (194919 images) from more than five hours to approximately 65 minutes. Additionally, the ability to achieve optimal performance while utilizing images with small dimensions (32 × 32 pixels) results in low memory and

TABLE 8. Proposed model compared with past distributed state-of-art methods.

Paper	Method	Image Size	Total image	Image Type	Classes	Accuracy
[54]	CoroNet	224×224	1251	Chest X-ray images	4	89.6%
[55]	CovXNet	128×128	1220	Chest X-ray images	4	90.3%
[56]	ResNetCOVID19	224×224	2634	Chest X-ray images	3	94.1%
[57]	COVID-FACT	256×256	307	CT-scan images	3	90.82%
[22]	Integration of ResNet and Location Attention Method	-	618	CT-scan images	3	86.7%
[51]	Proposed DML	488×408	2482	CT-scan images	2	98.91%
[58]	Hybrid metaheuristic and CNN algorithm	512×512	1097	CT-scan images	3	93.21%
[59]	Fine-tuned ResNet50V2	224×224	7593	CT-scan images	4	96.45%
Proposed Model	CTXNet	32×32	211655 X-ray and 194919 CT-scan	Chest X-ray and CT-scan images	4 of X-ray	95.37%
					3 of CT scan images	99.77%

storage requirements. This contributes to minimizing space and time complexity.

E. COMPARISON WITH PREVIOUS LITERATURE

The outcome of the study is compared with prior related literature. Table 8 represents a performance comparison of the proposed approach with some existing literature.

The table compares the proposed CTXNet model with past state-of-the-art methods. The proposed model, using a 32 × 32 image size, achieved remarkable accuracy rates of 95.37% for chest X-ray images and an impressive 99.77% for CT-scan images, with a total of 211,655 X-ray images and 194,919 CT-scan images. This is in contrast to other models such as Como Net and ConvXNet with accuracies of 89.6% and 90.3% respectively on chest X-ray images, and ResNet COVID-19 with a 94.1% accuracy on the same. COVID-FACT and the integration of ResNet and Location Attention Method dealt with CT-scan images, achieving accuracies of 90.82% and 86.7%. Other models like the Proposed DML and Hybrid metaheuristic and CNN algorithm showed accuracies of 98.91% and 93.21% on CT-scan images, while the fine-tuned Res Net 50V2 achieved 96.45%. In short, the proposed CYXNET model outperformed other models in terms of accuracy on both X-ray and CT-scan images.

Table 7 shows that TL models achieved a low accuracy from 43% to 77%, because image resolution is very important for preparing deep learning models. In the study of [60] authors conducted an analysis of the performance of widely recognized deep CNN models across various image resolutions from 32 × 32 pixels to 600 × 600 pixels. Reduced pixel size in images can lead to a loss of crucial information necessary for accurate classification by CNNs, ultimately decreasing accuracy. Table 8 shows that previous studies have shown accuracies of over 90%, that is because of the high-resolution images. Our proposed model CTXNET addresses the limitation of low-resolution images (32 × 32 pixels) by achieving high accuracy in terms of both CT scan and X ray dataset.

VIII. CONCLUSION

A lung disorder CAD system for two different modalities is presented in this work to categorize chest X-ray and CT scan images. The X-ray dataset used for the investigation had an insufficient and uneven quantity of images in distinct classes. The dataset was therefore balanced and the volume increased using the DCGAN data augmentation approach. Image pre-processing approaches were exploited to eradicate artifacts from the images. A vision transformer-based CTXNET model is proposed as it requires a shorter processing time and is trained on the CT-scan dataset. Ablation studies were conducted to assess and enhance the robustness of CTXNET. The model was also assessed with the X-ray dataset. Good performance was found for both datasets. 6 transfer learning models were tested and compared with the proposed CTXNET model based on accuracy and training duration using 32 × 32 pixel-sized images. Additionally, experiments were carried out by gradually decreasing the number of images of both datasets and training the model to assess the performance stability over image number. The proposed model performed remarkably well not only for the CT scan dataset, but also for the X-ray dataset, achieving a test accuracy of 99.77% for the CT scan dataset and 95.37% for the X-ray dataset, and requiring

only 10-12 and 40-42-seconds training time per epoch respectively. Using the same sized images with the other traditional models required 61-90 and 170-175 seconds per epoch while yielding comparatively poor accuracies ranging from 43% to 77% and 49% to 73% for chest X-ray and CT Scan datasets respectively. Furthermore, while the model was evaluated multiple times by decreasing the number of images, a consistency of performance is found which further validates the robustness of the framework. To further enhance the interpretability of the model's conclusions, the study also examined the usefulness of the Grad-CAM-based color visualization approach. This approach provides visual explanations that enable nuclear physicians to make quick and confident decisions based on the model's classifications. By combining accurate classification with explanatory visualizations, the proposed lung disorder CAD system holds promise for assisting medical professionals in diagnosing and treating lung disorders more effectively.

However, to address some limitations of the work, the whole work is performed on 2D CT scan data, although originally CT scans were 3D images. In future, the work could be carried out on 3D data. In addition, while our image pre-processing methods work well for this dataset, more research can be done on various image processing methods to handle noisy input photos, such as segmenting the various aspects of the chest images. It is also possible to assess the performance of our suggested model using real-time data. We could also look into graph and geometrical-based studies to comprehend the evolution of the chest disease.

FUNDING

This research received no external funding.

CONFLICTS OF INTEREST

The authors declare no potential conflicts of interests.

REFERENCES

[1] H. Wang et al., "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the global burden of disease study 2015," *Lancet*, vol. 388, no. 10053, pp. 1459–1544, Oct. 2016, doi: 10.1016/s0140-6736(16)31012-1.

[2] P. Yadav, N. Menon, V. Ravi, and S. Vishvanathan, "Lung-GANs: Unsupervised representation learning for lung disease classification using chest CT and X-ray images," *IEEE Trans. Eng. Manag.*, vol. 70, no. 8, pp. 2774–2786, Aug. 2023, doi: 10.1109/TEM.2021.3103334.

[3] L. H. Garland, "On the scientific evaluation of diagnostic procedures," *Radiology*, vol. 52, no. 3, pp. 309–328, Mar. 1949, doi: 10.1148/52.3.309.

[4] D. Shen and G. Wu. (2017). *Deep Learning in Medical Image Analysis*. Accessed: Oct. 20, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5479722/

[5] A. Arunachalam, V. Ravi, V. Acharya, and T. D. Pham, "Toward data-model-agnostic autonomous machine-generated data labeling and annotation platform: COVID-19 autoannotation use case," *IEEE Trans. Eng. Manag.*, vol. 70, no. 8, pp. 2695–2706, Aug. 2023, doi: 10.1109/TEM.2021.3094544.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[8] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9. Accessed: Oct. 20, 2022.

[10] D. Mukhtorov, M. Rakhmonova, S. Muksimova, and Y.-I. Cho, "Endoscopic image classification based on explainable deep learning," *Sensors*, vol. 23, no. 6, p. 3176, Mar. 2023, doi: 10.3390/s23063176.

[11] R. H. Abiyev and M. K. S. Ma'aitah, "Deep convolutional neural networks for chest diseases detection," *J. Healthcare Eng.*, vol. 2018, pp. 1–11, Aug. 2018, doi: 10.1155/2018/4168538.

[12] S. Varela-Santos and P. Melin, "Classification of X-ray images for pneumonia detection using texture features and neural networks," in *Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications* (Studies in Computational Intelligence), vol. 862. Cham, Germany: Springer, 2020, pp. 237–253, doi: 10.1007/978-3-030-35445-9_20.

[13] K. Wang, X. Zhang, S. Huang, and F. Chen, "Automatic detection of pneumonia in chest X-ray images using cooperative convolutional neural networks," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11858. Singapore: Springer, 2019, pp. 328–340, doi: 10.1007/978-3-030-31723-2_28.

[14] S. Thakur, Y. Goplani, S. Arora, R. Upadhyay, and G. Sharma, "Chest X-ray images based automated detection of pneumonia using transfer learning and CNN," in *Proc. Int. Conf. Artif. Intell. Appl.*, in Advances in Intelligent Systems and Computing, vol. 1164, 2021, pp. 329–335, doi: 10.1007/978-981-15-4992-2_31.

[15] A. Gupta, S. Gupta, and R. Katarya, "InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using chest X-ray," *Appl. Soft Comput.*, vol. 99, Feb. 2021, Art. no. 106859, doi: 10.1016/j.asoc.2020.106859.

[16] R. Jain, M. Gupta, S. Taneja, and D. J. Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images," *Appl. Intell.*, vol. 51, no. 3, pp. 1690–1700, Mar. 2021, doi: 10.1007/s10489-020-01902-1.

[17] M. Turkoglu, "COVIDetectioNet: COVID-19 diagnosis system based on X-ray images using features selected from pre-learned deep features ensemble," *Appl. Intell.*, vol. 51, no. 3, pp. 1213–1226, Mar. 2021, doi: 10.1007/s10489-020-01888-w.

[18] C. Ouchicha, O. Ammor, and M. Meknassi, "CVDNet: A novel deep learning architecture for detection of coronavirus (COVID-19) from chest X-ray images," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110245, doi: 10.1016/j.chaos.2020.110245.

[19] 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," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103792, doi: 10.1016/j.compbiomed.2020.103792.

[20] R. Ghavami Modegh et al., "Accurate and rapid diagnosis of COVID-19 pneumonia with batch effect removal of chest CT-scans and interpretable artificial intelligence," 2020, *arXiv:2011.11736*.

[21] D. Li, Z. Fu, and J. Xu, "Stacked-autoencoder-based model for COVID-19 diagnosis on CT images," *Int. J. Speech Technol.*, vol. 51, no. 5, pp. 2805–2817, May 2021, doi: 10.1007/s10489-020-02002-w.

[22] X. Xu et al., "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, Oct. 2020, doi: 10.1016/j.eng.2020.04.010.

[23] H. Mukherjee, S. Ghosh, A. Dhar, S. M. Obaidullah, K. C. Santosh, and K. Roy, "Deep neural network to detect COVID-19: One architecture for both CT scans and chest X-rays," *Int. J. Speech Technol.*, vol. 51, no. 5, pp. 2777–2789, May 2021, doi: 10.1007/s10489-020-01943-6.

[24] M. Y. Kamil, "A deep learning framework to detect COVID-19 disease via chest X-ray and CT scan images," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, p. 844, Feb. 2021, doi: 10.11591/ijece.v11i1.pp844-850.

[25] M. M. Ahsan, K. D. Gupta, M. M. Islam, S. Sen, M. L. Rahman, and M. S. Hossain, "COVID-19 symptoms detection based on NasNetMobile with explainable AI using various imaging modalities," *Mach. Learn. Knowl. Extraction*, vol. 2, no. 4, pp. 490–504, Oct. 2020, doi: 10.3390/make2040027.

[26] D. Dansana, R. Kumar, A. Bhattacharjee, D. J. Hemanth, D. Gupta, A. Khanna, and O. Castillo, "Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm," *Soft Comput.*, vol. 27, no. 5, pp. 2635–2643, Aug. 2020, doi: 10.1007/s00500-020-05275-y.

[27] A. Sedik, A. M. Iliyasu, B. A. El-Rahiem, M. E. A. Samea, A. Abdel-Raheem, M. Hammad, J. Peng, F. E. A. El-Samie, and A. A. A. El-Latif, "Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections," *Viruses*, vol. 12, no. 7, p. 769, Jul. 2020, doi: 10.3390/v12070769.

[28] *COVID-19 Radiography Database*. Accessed: Jun. 25, 2023. [Online]. Available: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[29] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image data augmentation for deep learning: A survey," 2022, *arXiv:2204.08610*.

[30] A. Eslambolchi, A. Maliglig, A. Gupta, and A. Gholamrezanezhad, "COVID-19 or non-COVID viral pneumonia: How to differentiate based on the radiologic findings?" *World J. Radiol.*, vol. 12, no. 12, pp. 289–301, Dec. 2020, doi: 10.4329/wjr.v12.i12.289.

[31] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 990–994, doi: 10.1109/ICASSP.2018.8461430.

[32] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–32.

[33] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. Valdés Hernández, J. Wardlaw, and D. Rueckert, "GAN augmentation: Augmenting training data using generative adversarial networks," 2018, *arXiv:1810.10863*.

[34] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On data augmentation for GAN training," *IEEE Trans. Image Process.*, vol. 30, pp. 1882–1897, 2021, doi: 10.1109/TIP.2021.3049346.

[35] F. Farahanipad, M. Rezaei, M. S. Nasr, F. Kamangar, and V. Athitsos, "A survey on GAN-based data augmentation for hand pose estimation problem," *Technologies*, vol. 10, no. 2, p. 43, Mar. 2022, doi: 10.3390/technologies10020043.

[36] L. Jin, F. Tan, and S. Jiang, "Generative adversarial network technologies and applications in computer vision," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–17, Aug. 2020, doi: 10.1155/2020/1459107.

[37] D. Wang, Y. Arzhaeva, L. Devnath, M. Qiao, S. Amirgholipour, Q. Liao, R. McBean, J. Hillhouse, S. Luo, D. Meredith, K. Newbigin, and D. Yates, "Automated pneumoconiosis detection on chest X-rays using cascaded learning with real and synthetic radiographs," in *Proc. Digit. Image Computing: Techn. Appl. (DICTA)*, Nov. 2020, pp. 1–6, doi: 10.1109/DICTA51227.2020.9363416.

[38] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020, doi: 10.1109/ACCESS.2020.2994762.

[39] N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and S. Elghamrawy, "Detection of coronavirus (COVID-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest X-ray dataset," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.*, in Lecture Notes on Data Engineering and Communications Technologies, vol. 152. Cham, Switzerland: Springer, 2023, pp. 234–247, doi: 10.1007/978-3-031-20601-6_22.

[40] L. Devnath, S. Luo, P. Summons, and D. Wang, "An accurate black lung detection using transfer learning based on deep neural networks," in *Proc. Int. Conf. Image Vis. Comput. New Zealand*, Dec. 2019, pp. 1–6, doi: 10.1109/IVCNZ48456.2019.8960961.

[41] P. Ghosh, S. Azam, R. Quadir, A. Karim, F. M. J. M. Shamrat, S. K. Bhowmik, M. Jonkman, K. M. Hasib, and K. Ahmed, "SkinNet-16: A deep learning approach to identify benign and malignant skin lesions," *Frontiers Oncol.*, vol. 12, Aug. 2022, Art. no. 931141, doi: 10.3389/fonc.2022.931141.

[42] T. M. Breuel, "Efficient binary and run length morphology and its application to document image processing," 2007, *arXiv:0712.0121*.

[43] P. Dhar, "A method to detect breast cancer based on morphological operation," *Int. J. Educ. Manage. Eng.*, vol. 11, no. 2, pp. 25–31, Apr. 2021.

[44] A. R. Beeravolu, S. Azam, M. Jonkman, B. Shanmugam, K. Kannoorpatti, and A. Anwar, "Preprocessing of breast cancer images to create datasets for deep-CNN," *IEEE Access*, vol. 9, pp. 33438–33463, 2021.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, Sep. 2014, pp. 1–14.

[46] X. Huang, N. Bi, and J. Tan, "Visual transformer-based models: A survey," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 13364. Cham, Germany: Springer, 2022, pp. 295–305, doi: 10.1007/978-3-031-09282-4_25.

[47] K. Islam, "Recent advances in vision transformer: A survey and outlook of recent work," 2022, *arXiv:2203.01536*.

[48] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[49] S. Hossain, S. Azam, S. Montaha, A. Karim, S. S. Chowa, C. Mondol, M. Z. Hasan, and M. Jonkman, "Automated breast tumor ultrasound image segmentation with hybrid UNet and classification using fine-tuned CNN model," *Heliyon*, vol. 9, no. 11, pp. 1–31, 2023, doi: 0.1016/j.heliyon.2023.e21369.

[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[51] K. Gupta and V. Bajaj, "Deep learning models-based CT-scan image classification for automated screening of COVID-19," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104268, doi: 10.1016/j.bspc.2022.104268.

[52] M. A. Khan, M. Azhar, K. Ibrar, A. Alqahtani, S. Alsubai, A. Binbusayyis, Y. J. Kim, and B. Chang, "COVID-19 classification from chest X-ray images: A framework of deep explainable artificial intelligence," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–14, Jul. 2022, doi: 10.1155/2022/4254631.

[53] S. Montaha, S. Azam, A. K. M. R. H. Rafid, M. Z. Hasan, A. Karim, K. M. Hasib, S. K. Patel, M. Jonkman, and Z. I. Mannan, "MNet-10: A robust shallow convolutional neural network model performing ablation study on medical images assessing the effectiveness of applying optimal data augmentation technique," *Frontiers Med.*, vol. 9, Aug. 2022, Art. no. 924979, doi: 10.3389/fmed.2022.924979.

[54] 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," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105581, doi: 10.1016/j.cmpb.2020.105581.

[55] T. Mahmud, M. A. Rahman, and S. A. Fattah, "CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103869, doi: 10.1016/j.compbiomed.2020.103869.

[56] M. Arsenovic, S. Sladojevic, S. Orcic, A. Anderla, and M. Sladojevic, "Detection of COVID-19 cases by utilizing deep learning algorithms on X-ray images," in *Proc. 18th Int. Sci. Conf. Ind. Syst. Ind. Innov. Digital Age*, Jan. 2020, pp. 1–8.

[57] S. Heidarian, P. Afshar, N. Enshaei, F. Naderkhani, M. J. Rafiee, F. B. Fard, K. Samimi, S. F. Atashzar, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "COVID-FACT: A fully-automated capsule network-based framework for identification of COVID-19 cases from chest CT scans," *Frontiers Artif. Intell.*, vol. 4, p. 65, May 2021, doi: 10.3389/frai.2021.598932.

[58] T. I. A. Mohamed, O. N. Oyelade, and A. E. Ezugwu, "Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm," *PLoS ONE*, vol. 18, no. 8, Aug. 2023, Art. no. e0285796, doi: 10.1371/journal.pone.0285796.

[59] K. U. Ahamed, M. Islam, A. Uddin, A. Akhter, B. K. Paul, M. A. Yousuf, S. Uddin, J. M. W. Quinn, and M. A. Moni, "A deep learning approach using effective preprocessing techniques to detect COVID-19 from chest CT-scan and X-ray images," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 105014, doi: 10.1016/j.compbiomed.2021.105014.

[60] C. F. Sabottke and B. M. Spieler, "The effect of image resolution on deep learning in radiography," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190015, doi: 10.1148/ryai.2019190015.