IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 July 2024, accepted 8 August 2024, date of publication 15 August 2024, date of current version 29 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444601

RESEARCH ARTICLE

RAIDER: Rapid AI Diagnosis at Edge Using Ensemble Models for Radiology

ISHAN ARYENDU, (Graduate Student Member, IEEE), AND YING WANG, (Member, IEEE)
School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, USA
Corresponding author: Ying Wang (ywang6@stevens.edu)

ABSTRACT Chest X-rays have played an indispensable part in medical diagnosis for several decades. However, there is a scarcity of experts who can interpret these images to diagnose critical illnesses, which can lead to preventable fatalities. This paper introduces a novel Rapid AI Diagnosis at Edge using Ensemble Models for Radiology (RAIDER) designed to leverage the advantages of cross-geolocation meta-learning models. We can generate local machine learning models at individual locations and distribute them across other locations for diagnosing diseases at the edge or on-premises if required before they become worldwide pandemics, significantly enhancing the rapid or near-real-time identification of fast-spreading respiratory diseases through online learning. This novel approach allows for geo-distributed multi-fold model training, harnessing the unique strengths of diverse geographical data sources to improve diagnostic accuracy and speed by leveraging edge computing. Using the existing Convolutional Neural Network (CNN) models and distributed training at the edge, we can enhance the accuracy and cost-effectiveness of diagnosis. The proposed architecture allows for distributed training and independently verified performance metrics on the MIMIC-CXR and COVIDGR chest X-ray datasets with accuracy, sensitivity, specificity, F1-score and AUC of **97.80%**, **97.06%**, **98.48%**, **96.51%**, and **0.9739**, respectively. Our proposed RAIDER architecture marks the first implementation of a collaborative framework that facilitates seamless interaction across different geographic locations and edge computing, enabling a more effective and efficient response to emerging health threats.

INDEX TERMS Chest x-ray, medical image diagnosis, edge computing, ensemble learning, meta-learning.

I. INTRODUCTION

We live in a tender world with a struggling health infrastructure to care for the ever-growing population on our planet. The fragile nature of our society was on full display at the onset of the rapid spread of COVID-19, which caused a global pandemic a few years ago [1]. The hospitals were swamped with hundreds of thousands of patients, requiring swift diagnosis to prevent overcrowding and further spread of diseases. Since then, several COVID-19 variants like Beta, Delta, and Omicron have caused global public health concerns in the following years due to their resistance to the existing vaccines [2] as they were less susceptible to neutralization by antibodies generated by previous infection or existing vaccination. In light of these events, the critical infrastructure supporting rapid diagnosis has attracted significant attention [3], [4]. While real-time reverse transcription-polymerase chain reaction (RT-PCR) is the most widely used method, it takes a long time [5], [6], and has a sensitivity of only 60%-70%, meaning that it is crucial to have commercial kits available in order to test positive in about 30% of situations when the test result is negative [7]. Although there are several alternatives for diagnosing patients with COVID-19, chest X-ray images have particularly proven useful in assessing the severity of the disease. Several research studies have explored the effectiveness and the potential of using machine learning models for interpreting these images. Interpreting a chest radiograph can present difficulties because of the superimposed anatomical structures along the projection path. This phenomenon may pose significant challenges in terms of identifying anomalies in specific areas like a nodule

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

posterior to the heart on a frontal CXR [8], detecting minute or subtle abnormalities, or precisely differentiating between various pathological patterns. For these reasons, analysis of CXR pictures by radiologists is often subject to considerable interobserver variability. Several deep learning models have been presented that detect COVID-19 with high accuracy [9], [10]. Nevertheless, these models are enormous, complex to scale, and expensive to train. They also require large datasets of images to detect subtle differences in the medical images, which need more nuanced image enhancements and complex models to detect these features as the number of diseases to classify grows. When a new COVID-19 variant emerges, the associated symptoms and protocols for diagnosing the disease via chest X-rays may change. Additionally, several other diseases also rely on chest X-rays for diagnosis [11]. In addition, the current supervised training of deep neural networks on medical image diagnosis relies heavily on large pools of labeled data, which is scarce, inflexible, expensive, and limited by types of annotations. Since chest X-rays are private and confidential data, they can only be shared with other researchers with proper anonymization and approval. However, anonymizing and using the data on-site to train the classification models is more straightforward. As conventional computer-aided diagnosis systems face significant hurdles in adapting to the rapid spreading of new strains of this disease, a more flexible and decentralized approach is needed to keep pace with the rapid spread of such infectious and life-threatening diseases. Using such an approach, we can standardize the image preprocessing steps and share the machine learning models without sharing the patients' private data. Furthermore, the approach would be scalable and enable collaboration across different geographic locations to quickly detect novel diseases and their variants.

II. BACKGROUND AND RELATED WORK

Researchers worldwide have used Chest X-ray images to assess and diagnose several health issues in patients, including assessing the condition of the lungs, heart-related problems, detecting the size and outline of the heart, blood vessels, calcium deposits, fractures, postoperative changes, the presence of a pacemaker, defibrillator or catheter, and several combinations of these. Several deep learning models trained on publicly available labeled chest X-ray datasets are available for detecting these illnesses. For example, CheXpert can automatically detect the presence of 13 diseases in radiology reports [12], including Enlarged Cardiom, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Fracture, Support Devices, and more. Similarly, CheXNet can also detect 14 diseases but outperforms CheXpert in terms of accuracy [13]. These models demonstrate the potential of using deep learning models in detecting pneumonia and other pathological diseases from chest X-rays by achieving performance levels that exceeded or matched those of practicing radiologists. Similar accuracy (81%) was shown by DeepCOVID-XR even when compared to the consensus of all five radiologists. Interestingly, DeepCOVID-XR demonstrated a notably greater specificity (92%) compared to two radiologists (75%, $P < 0.001$; 84%, $P = 0.009$) and a significantly higher sensitivity (71%) compared to a single radiologist (60%, $P < 0.001$). However, with a P-value of 0.13 coupled with the difference between DeepCOVID-XR's AUC of 0.88 and the consensus AUC of 0.85 was not statistically significant [14]. The disparity in performance throughout the training and testing phase has to be more persuasive to the radiologists, even though the performance was better than that of individual radiologists. Other studies have also found the CNN models to be on par with the assessment of the radiologist [15]. Although the design and training of these models become more complex as we introduce additional classes of diseases for classification, they offer a promising glimpse into the potential for deploying CNNs to identify novel diseases in real time. To benefit from such an approach, we require a model that balances efficiency with accurate classification capability, ensuring it is lightweight enough for time-sensitive applications yet sufficiently robust enough to classify diseases accurately.

A. LIGHT-WEIGHT PREDICTION MODELS

One of the popular methods of training a CNN model is extracting knowledge from one domain and transferring it to another through transfer learning. Pre-training is one of the transfer learning strategies that has been frequently applied in CXR analysis. The pre-training strategy starts out with training the network architecture on a large dataset for a different task. The trained weights obtained from this phase are then utilized as an initialization for the subsequent jobs to fine-tune them. Either all layers or just the last layer may be retrained, depending on the dataset of the target domain. Because important low-level features are acquired from the source data, this method enables neural networks to be trained for new tasks using comparatively smaller datasets. For the purpose of classifying real images, pre-training on the ImageNet dataset has been demonstrated to be advantageous [16]. Additionally, pre-trained architectures can be used for feature extraction in conjunction with more conventional techniques like random forests or support vector machines. Moreover, there have been many comprehensive studies about using lightweight convolutional neural networks (CNNs) to detect diseases using chest X-rays. Lightweight models from the SqueezeNet, VGG, and MobileNet model families have been determined to classify multiple classes of diseases with minor tweaks to their architectures [17].

B. DATA ACQUISITION DIVERSITY

In most previous research, researchers have trained their supervised deep neural networks (DNN) on a large pool of labeled data. However, in medical imaging, labeled data is scarce due to privacy concerns, and the manual annotation by professional radiologists and physicians requires tedious, time-consuming effort. Although some large labeled datasets

are available, they can be severely imbalanced by over-representing common problems and under-representing rare conditions. Moreover, the resolution of differential diagnosis in these diseases is generally low, so the fine detection and classification for diagnosis among images is not feasible. This imbalance can lead to poor performance of DNNs on rare diseases. As expected, we came across several datasets that had gathered ChestX-ray pictures of various health conditions, including MIMIC-CXR, BIMCV, COVIDDSL, and COVIDGR, consisting of images of various resolutions and quantities. The National Institutes of Health (USA) gathered 112,120 pictures of 30,805 patients to create ChestX-ray14 [18]. The images are grayscale 1024 × 1024 pixels with 8-bit labels from radiological reports that show 14 different disease kinds. The MIMIC-CXR dataset comprises 371,920 pictures gathered from 64,588 patients [19] who were admitted to Beth Israel Deaconess Medical Center's emergency department between 2011 and 2016. This dataset was likewise labeled from radiology records using the same rule-based labeler method as CheXpert. Chest X-rays and CT scans are included in the BIMCV COVID-19 dataset, provided by the Valencian Region Medical ImageBank in 2020 [20]. The dataset includes 3293 16-bit original resolution images from 1305 COVID-19 patients. The HM Hospitales group in Spain released the COVID-19 dataset known as COVIDDSL [21]. It contains comprehensive laboratory test results, vital signs, and CXR images for 1725 individuals. Confirmed COVID-19-positive status is reported for every individual. By related RT-PCR results acquired in less than 24 hours, half of the 852 PA CXR pictures in the COVIDGR dataset are classified as COVID-19 positive [18]. The degree of severity of positive cases is shown in this dataset, which was gathered from Hospital Universitario Clínico San Cecilio in Granada, Spain.

Despite the availability of numerous public chest X-ray datasets, preparing the data for use with proposed models remains challenging due to issues with annotations, labeling consistency, and varying data formats. Most Datasets use Natural Language Processing (NLP) to generate labels for each image. Even though it is a fast and inexpensive labeling method, it results in inaccuracies in labeling [12], [22]. There might be several causes for such inaccuracies, as some visible abnormalities may not be mentioned in the radiology report, depending on the context in which it was acquired. The NLP algorithm can be erroneous, interpreting some negative statements as positive, failing to identify acronyms, and many more. Furthermore, many findings on CXR are subtle or doubtful, leading to disagreements even among expert observers [23]. The CheXpert dataset mentions no labels or uncertain levels to account for this uncertainty. One particular cause for concern with NLP labels is the issue of systematic or structured mislabeling, where a disease is consistently mislabeled in the same way. This example occurs in the ChestX-ray14 dataset, where subcutaneous emphysema is frequently identified as pulmonary emphysema [24], [25].

However, we can use this fragmentation to our advantage. Since there are multiple specialized clinics for the treatment of specific diseases, we can get the labeled data from the experts in these institutions to train the CNN models precisely to diagnose these diseases. Nonetheless, there needs to be more clarity between the current deep learning-based methods and the medical demands that require the detection of subtle differences like which stage or at what scale the disease has spread through the medical images. For example, pulmonary edema is a fluid buildup in the lungs and is one of the most direct symptoms of Chronic Heart Failure (CHF) [26], where the heart cannot pump blood effectively. CHF patients have extremely heterogeneous responses to treatment [27] and respond differently to the same medications and interventions. This makes it difficult for clinicians to come up with effective treatment plans. Assessing the severity of pulmonary edema will enable clinicians to make better treatment plans based on prior patient responses, where deep learning methods are instrumental. It will facilitate clinical research studies that require quantitative phenotyping of the patient's status [28]. However, quantifying pulmonary edema is a highly challenging task. The grading of pulmonary edema severity relies on much more subtle image findings than detecting pathologies in chest X-ray images [13], [29]. To fill the gap between the current deep learning-based methods and the medical demands of detecting subtle differences in medical images, we propose a distributed, scalable system that uses image processing and standardizing techniques to enable user-defined targets for specific specialty areas. Various models are effective in the classification of specific diseases but not others. Therefore, ensembling methods for combining the predictions of these models can help us create a unifying solution for diagnosing these diseases.

C. ENSEMBLE LEARNING

Using ensemble learning, we can combine the benefits of several baseline models to build a better model than its constituents [30] with the added benefit of reducing the probability of overfitting. This technique has been successfully applied to several fields, including character recognition [31] and sentiment analysis [32]. Various ensembling strategies differ in how distinct baseline models are learned and blended. The most commonly utilized ensemble approaches are averaging, bagging, random forest, stacking, meta-learning, and boosting [33], [34], [35]. Nevertheless, most of these efforts apply average voting mechanism baseline deep-learning models. A majority voting classifier [36] for detecting pneumonia and COVID-19 performed admirably in classifying these diseases. The transfer learning approach combined with ensembling gives us satisfactory performance. However, as we will see in the results section, combining baseline learners using maximum voting is only sometimes the best strategy, as the ensembling process using average voting techniques is skewed toward weak baseline learners. Another popular technique for ensembling is using stacking.

In CovXNet [37] a meta-learner was used to stack the output of the models to give a respectable final prediction. However, the shallow neural network does not clearly show us the bias of the neural network towards any specific individual classification model. Therefore, we need a more interpretable model that can convince medical practitioners about these models' reliability and unbiased predictions. Pruning can help us reduce the complexity and eliminate the bias [38] in our ensembled model. However, pruning is computationally intensive and is not ideal for providing the results in a short time, voiding the objective of achieving large-scale diagnosis at a faster rate. Some other ensembling models, like DeepCOVID-XR [14], have used the weighted averages of the individual models to get decent results. However, the weights assigned to each model are subjective, and there is no definitive way to arrive at the optimal solution on the Pareto boundary, as each point on this boundary gives us the same value for the objective function. It is also possible that we might end up with a biased model which is impractical for sensitive medical diagnosis applications. Although numerous approaches for merging baseline learners can be used for group deep learning, these approaches have certain drawbacks concerning generalization and training challenges.

D. CONTRIBUTION

Our study presents an integrated system tailored to enhance diagnostic processes for respiratory diseases through detailed analysis of chest X-rays at the edge. The system's architecture is designed to balance robustness with efficiency. Our main contribution can be summarized as follows:

- Multiple light-weight prediction models are specifically optimized to differentiate between various respiratory conditions by examining the distinct features within the images. Notably, our system architecture permits the separation of various far-edge, near-edge, and on-premises predictive components, thereby enhancing flexibility and allowing for extensive feature extraction and model training without the constraints of immediate output.
- We proposed an innovative ensembling strategy incorporating meta-learning and deftly consolidating the predictive outputs. The flexibility of our architecture shines in this component, enabling real-time ensemble learning to provide prompt and request-oriented classification while preserving the integrity of the in-depth training previously conducted. This dual-structure ensembling is instrumental in delivering swift and precise final diagnoses.
- Our methodology deliberately utilizes a limited but representative subset of images for each disease class, reflecting the realistic scenarios of data limitation typically encountered during novel disease emergence at a specialty center. The classification models serve as the backbone for disease identification through a decision tree for ensembling and ascertaining definitive disease classifications.
- The distinctive separation of our system's prediction and ensembling components fosters a flexible, decentralized, and scalable architecture. This architecture balances the need for rapid and rigorous feature extraction training with the demand for efficient classification on demand.

The remainder of the paper begins with an overview of the system architecture, followed by a description of our dataset. Subsequently, we will discuss the specifics of the image processing techniques, the ensembling model we utilize, the experimental setup, the results obtained, and a discussion of the results. The conclusion will follow at the end of this paper.

III. SYSTEM DESIGN AND METHODOLOGY

A. SYSTEM ARCHITECTURE

In Fig. 1, we describe the system diagram of our proposed solution. We collect several chest X-rays regularly, which the specialists label according to the patient's medical condition at the hospitals. These labeled CXR images are then processed on-premises or can be sent to the edge servers, where these images are processed according to the process described in Fig. 2. If we detect a new class of disease that is not in our database, we send notification alerts to all the physicians in our network about a potential outbreak of a novel disease so that they can take preventive measures to contain the spread of the disease. Furthermore, multiple neural networks are trained at the edge servers to classify new diseases, which will be stored securely on a cloud server that hosts the patients' medical records. We can encrypt sensitive personal information and obfuscate it to preserve the privacy and confidentiality of the information stored on the servers. This process will happen in the background as a part of the scheduled batch processes. Suppose a particular clinic, lab, or physician needs access to these models. Depending on their needs, we can give them individual models or ensembled models, which employ a lightweight meta-learning approach and can be used on mobile devices to categorize diseases based on available CXR images. This will act as getting a second opinion from experts. We will periodically fetch the meta-learning model from our distributed devices to keep the local model in sync with the latest one available on the servers. This is an overview of how we plan to process the CXRs generated at the hospital and use machine learning models to assist the physicians in diagnosing the patients' health conditions. The details about the crucial components of the system architecture are described below:

- **Distributed Learning:** We have several sites, which can be specialty centers for diseases or research labs, working on detecting several respiratory diseases. Each site maintains its own database of anonymized chest X-rays, which will be synced with a decentralized database, with provisions for redundancy, in the cloud. Anyone authorized to put the images on these servers

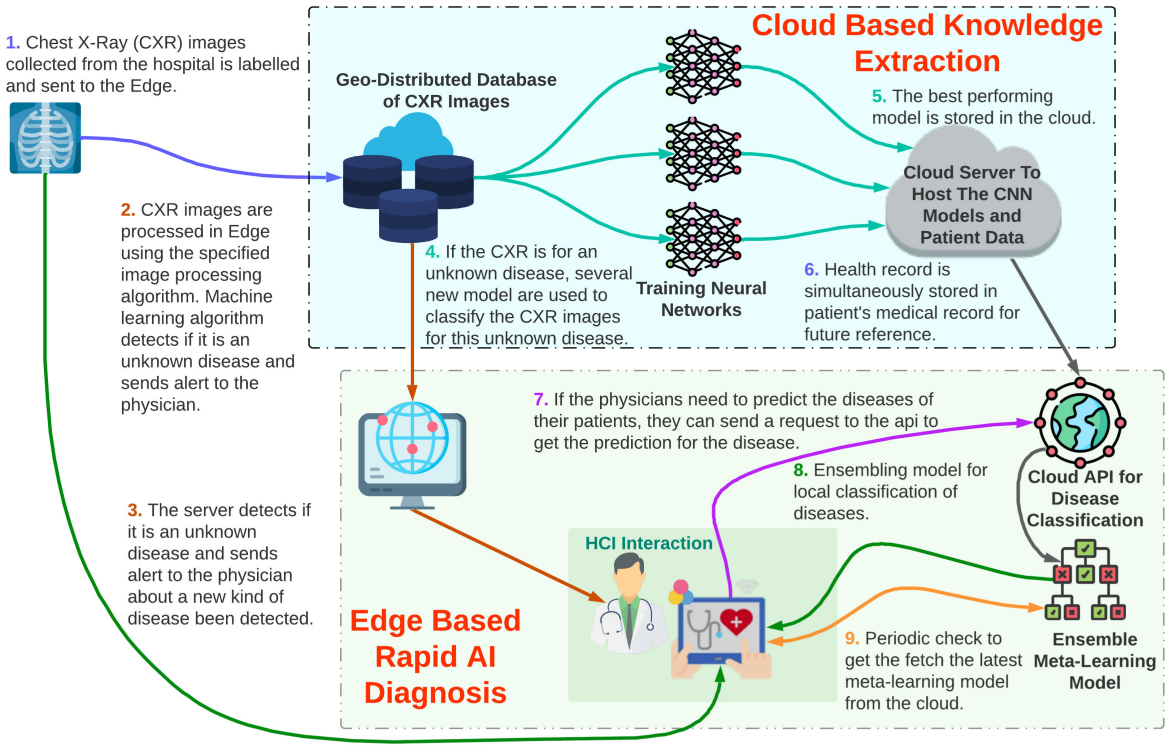


FIGURE 1. Collaborative Ensembling Architecture: After preprocessing of CXR images at the edge, these images are analyzed using a local ensembling model that combines predictions from multiple models. These models are periodically synchronized with the latest meta-learning models from the cloud to keep them fresh. Physicians can interactively transmit CXR images and obtain disease classification results through this local model, facilitating rapid diagnosis in clinical communication (HCI Communication). When an unknown disease is detected, images are sent to the cloud for further analysis and new modeling, ensuring the system adapts to emerging health risks.

will specify the label associated with the diagnosis of the image and will store the individual's digital signature along with the image for future audits.

- **Training Servers:** We will have several commodity hardware that will serve as the distributed training servers and will run multiple models and use several metrics as their benchmark to choose the best model for classifying the disease. Using the image data gathered from the patients, these models will be hosted on distributed servers with restricted access to prevent unauthorized access and manipulation. The physicians will play a critical role in feature extraction, model selection, and optimization processes by training on the server to create the model.
- **Ensemble System:** This component combines the results from all sites and can use several approaches, such as voting, stacking, meta-learning, etc, to aggregate predictions from multiple models and reduce computational cost.
- **HCI Interaction System:** The "ENSEMBLE" model can run locally on the handheld device to get the final prediction, or we can get the predictions from one of the edge servers. This approach allows for diverse real-time consultation and can mitigate biases inherent to individual models and radiologists using knowledge from different parts of the world.

TABLE 1. Resolution of images in the dataset.

Image Class	Width		Height	
	Min	Max	Min	Max
Normal	1040	2628	650	2628
COVID-19	240	3520	237	4095
Viral Pneumonia	384	2304	127	230

B. DATASET DESCRIPTION

As seen in Table 1, the MIMIC-CXR and COVIDGR datasets we are working with have images with several resolutions, which might throw off the classification models during training. Therefore, the standardization of the image resolutions is required to mitigate the effects of the image sizes and pixel counts on our models. Subsequently, integrating a standard range of pixels will also make the dataset of images widely accessible to researchers who will require minimal preprocessing in terms of standardization and improve collaborative research. We randomly sampled 595 COVID-19, 618 Viral Pneumonia, and 625 images for Normal cases. We resized the images to 256 × 256 pixels. Furthermore, we normalized these pixels to a value between 0-1 for more efficient training of the neural networks. While CNN-based models anticipate a 3-channel picture, radiography images are usually 1-channel images. In order to prepare our single-channel photos for use with models intended

for three-channel RGB inputs, we replicate the grey-scale information in each image channel.

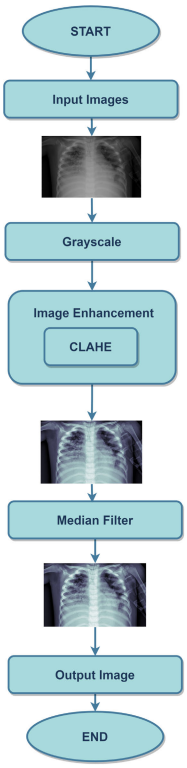


FIGURE 2. Flowchart showing the step-by-step process of image enhancement. Starting with the grayscale conversion of input images, we apply CLAHE and then further refine the images using a median filter, resulting in the final output image.

C. IMAGE PREPROCESSING

X-rays pass through an object and interact with the photographic emulsion on a film to create a radiographic image [39]. The chemicals in the capturing film react with the X-rays and turn dark. The discoloration depends upon the amount of X-rays hitting the film, which depends on the object's density. So, these images have superimposed black, white, and grey shadows. The perceived density of the adjacent structures may change depending on their contrast [40]. Therefore, complex processing is required to reach the diagnosis from these images. Several studies have been conducted on enhancing the images to improve their clarity [41]. We will use the approach described in the flowchart shown in Fig. 2.

1) GRAYSCALE CONVERSION

The images obtained from various sources might have more than one-channel of color. Therefore, we'll convert them into grayscale to better control the training dataset. This can be done using the formula [42] below:

$$Grayscale = (0.299 \times R) + (0.587 \times G) + (0.114 \times B) \tag{1}$$

where,

- R = red channel matrix value
- G = green channel matrix value
- B = blue channel matrix value

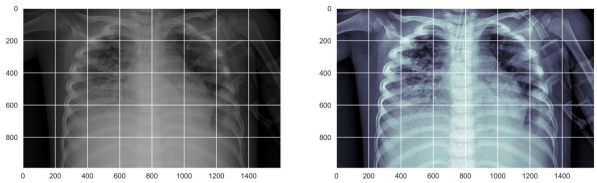
2) CONTRAST STRETCHING

This technique stretches the contrast by expanding the dynamic range of the image's intensity value. Using linear scaling to apply the image's pixel value is possible. Finding the image's minimum and maximum values is required in order to normalize or contrast-extend the image. The image boundary is defined by these minimum and maximum values. The lowest limit of this suggested method is an image with an 8-bit gray level, while the image density has a minimum value of 0 and a maximum value of 255. The equation is given by:

$$g(x, y) = \frac{f(x, y) - min}{max - min} \times 255 \tag{2}$$

where,

- $g(x, y)$ = matrix of the resulting image
- $f(x, y)$ = original image matrix value



(a) Original Image (b) Image after CLAHE

FIGURE 3. X-ray images of COVID-19 before and after applying CLAHE.

3) HISTOGRAM EQUALIZATION

Histogram equalization (HE) is widely used to flatten the gray distribution of images by adjusting the values of corresponding image pixels to yield higher contrast. The flattened histogram distribution function is then applied to the image. The output of the process gives us the intensity and scale for the pixel value at the gray level. The goal of this HE is to create an evenly distributed histogram where input images with different intensity levels will produce output images with the same contrast level [43] at a global level. Adaptive histogram equalization (AHE) has been used as a popular method to enhance the contrast of images. Using several histograms, the image's luminance is calculated for different sections of the image, which results in better edge visualization and increased local contrast. However, the noise might be amplified in homogeneous regions. This can be prevented if we use contrast-limited adaptive histogram equalization (CLAHE) [44], [45] for local contrast enhancement.

The advancement of HE and AHE into CLAHE, a contrast enhancement technique, makes CLAHE a superior alternative to HE and AHE. Because CLAHE restricts contrast to local blocks, it uses the maximum value on the local grids to clip and return the gray values. Images in CLAHE are split into tiles for better manageable regional sections, which prevents

noisy images. The equation below describes the computation of CLAHE,

$$p = (p_{max} - p_{min}) \cdot P(f) + p_{min} \tag{3}$$

where the new pixel value p is obtained by applying a formula using the maximum and minimum pixel values of the image p_{max} and p_{min} and the cumulative probability distribution function $P(f)$. In other words, the new pixel value is computed by scaling the cumulative probability distribution function to the range between p_{max} and p_{min}, and then adding p_{min} to the result. Fig. 3 compares the effects of applying CLAHE on the original image.

4) MEDIAN FILTER

To preserve image information, noise in the image can be filtered or reduced using a non-linear technique called median filtering. Using values as medians will change or replace values on images. The gray level replaces the value with the median, which is subsequently stored in place of the noise value. Suppose the highest is given by the max and the minimum is given by the min, and then the average represents the gray level. Additionally, the median filter's manipulation center, y (m, n) [46], is given as:

$$y[m, n] = median\{x[i, j], (i, j) \in \omega\} \tag{4}$$

where,

- $y[m, n]$ = matrix of results labeled as y with m, n as a rows and columns.
- $\{x[i, j], (i, j) \in \omega\}$ = matrix value of the image being processed or the corresponding elements sorted.

Once the neighboring pixel value is determined, the median filter compares the current value to the value of the neighboring pixel values. At the intermediate value, the pixel value will be changed based on the count of neighboring pixels. An average of two middle values determines the median value [47].

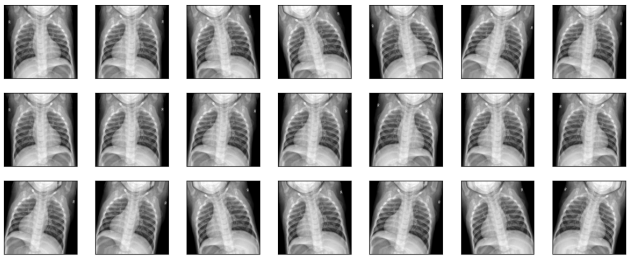


FIGURE 4. Depiction of 21 augmented X-Ray images for machine learning training. These augmented images, which include rotation, scaling, flipping, and cropping, improve the robustness of machine-learning models during training.

D. IMAGE AUGMENTATION

Data augmentation significantly expands the variety of the existing data sets without adding new data. Techniques such as cropping, padding, and horizontal flipping are commonly used to train large neural networks. It can improve accuracy and model training by creating multiple copies of the same image from different methods, such as rotation, rotation, sharpening, and many other methods. In previous studies [48], the authors have used a deep convolutional neural network (CNN) trained on the CIFAR-10 dataset to perform geometric transformations. The accuracy of the CNN was 91.8%, which is much higher than the 89.6% accuracy obtained without geometrically implemented transformation.

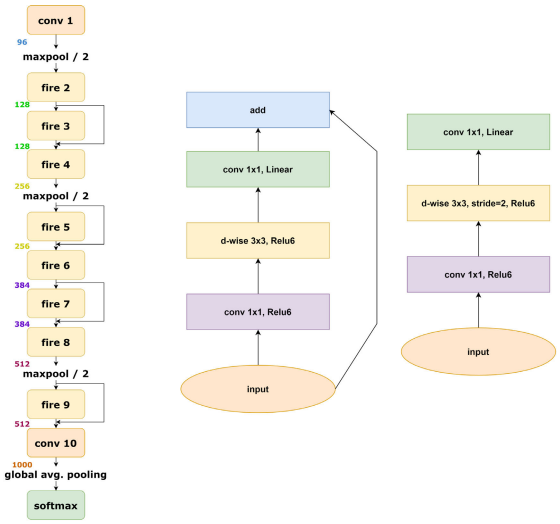
As can be seen from their studies, geometric transformation improves the performance of the image classification model by introducing more diversity in the training process and enabling the model to cope with changes in the orientation, shape, and location of segmented objects. Which in turn significantly improves the performance of image classification models. By providing more diversity in the training set, the geometric variations contribute to the model having more complex features that are less sensitive to the direction of changes in distributed features, size, and location, thus improving the accuracy and generalizability of unobserved test data. Fig. 4, shows the effects of applying image augmentation on a sample of images that would be used to train the model.

E. ENSEMBLING ON SELECTED LIGHTWEIGHT PREDICTION MODELS

Following a comparative analysis of various lightweight prediction models in the background research, we have selected MobileNetV2 and SqueezeNet for our framework because of the fewer parameters to train and their wider classification capabilities. MobileNetV2 stands out for its exceptional efficiency and accuracy in mobile and embedded vision applications, benefiting from a streamlined architecture that minimizes computational requirements while maintaining high performance [49]. SqueezeNet offers remarkable model compactness through the use of squeeze and expansion layers, achieving AlexNet-level accuracy with a fraction of the parameters, making it highly suitable for environments with strict memory limitations [50]. The proposed ensembling [51], [52] framework can also be applied to other predictive models, enhancing its versatility and applicability across different domains. The architecture of MobileNetV2 and SqueezeNet with bypass are shown in Fig. 5.a and Fig. 5.b, respectively that would be a good fit for the type of problem we are trying to solve. However, the data to train these models are sourced from several datasets. Hence, it becomes imperative that we preprocess the data before training the models.

An aggregation function A combines the results of n baseline classifiers $c_1, c_2, \ldots, c_n$ to predict the final output. Suppose we have a dataset of size a and features of dimension b, $D = \{(x_i, y_i)\}_{i=1}^{a}$, where $1 \leq i \leq a$ and $x_i \in \mathbb{R}^b$. The prediction of the output based on this ensemble method is given by the equation below:

$$y_i = \Omega(x_i) = A(c_1, c_2, \ldots, c_n). \tag{5}$$



(a) Architecture of SqueezeNet model (b) Architecture of MobilenetV2 model

FIGURE 5. Achitecture diagrams for SqueezeNet and MobileNetV2 models.

We will use average voting [53] and meta-learning [54], two popular ensemble learning techniques in this study. The concept behind average voting is that predictions are taken from several models, and the final prediction is determined by averaging these predictions [55]. As the function of the equation below illustrates, the arithmetic mean—the sum of the forecasts divided by the total number of guesses—is used to get the average prediction.

$$y^* = argmax\left(\frac{1}{n_j}\sum_{j=1}^{m} w_{i,j}\right) \tag{6}$$

Consider a scenario in which the ensemble consisted of the following three classifiers: $c_1(x) = [0.63, 0.17, 0.20]$, $c_2(x) = [0.28, 0.51, 0.21]$, and $c_3(x) = [0.61, 0.29, 0.10]$. Using the average weight approach, the following would be the mean prediction $y_0 = [0.63, 0.28, 0.61]/3 = 0.506$ for class 0 and a prediction of $y_1 = [0.29 + 0.17 + 0.51]/3 = 0.323$ for the class 1 and $y_2 = [0.20 + 0.21 + 0.10]/3 = 0.17$ for class 2.

The other ensemble learning method is learning from learners, popularly known as meta-learning [54], [56], which depends on prior experience with previous classification models. By altering some parts of the learning algorithm in response to experimental findings, we can enhance its performance and outcomes. In contrast to conventional machine-learning models, the meta-learning approach uses multiple learning stages, with each stage inducing its output as an input to the meta-learner, which produces the final output [57]. Moreover, meta-learning helps learning algorithms better adapt to changing circumstances, expedites learning processes by lowering the number of tests needed, and optimizes hyperparameters to provide ideal outcomes. Additionally, this approach offers the chance to address several deep learning challenges, such as generalization, computational complexity, and data size [58]. Our approach is detailed in the flowchart in Fig. 6, where we start with individual model predictions, stack them, and convert the target labels to a 1-D array. The meta-learning model is then trained and evaluated on the validation dataset, which yields the final output.

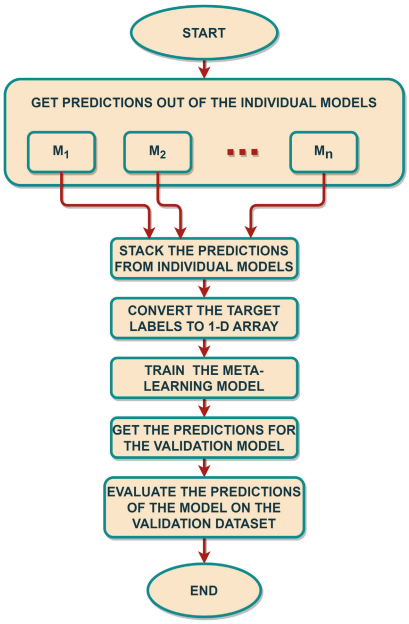


FIGURE 6. Meta-learning model flowchart illustrating the step-by-step process starting with individual model predictions, stacking these predictions, and converting the target labels to a 1-D array. The meta-learning model is then trained and evaluated on the validation dataset, resulting in the final output.

IV. EXPERIMENTAL SETUP

We have used the TensorFlow library to build our neural networks. For our study, we used a Windows server 2022 machine functioning as the far-edge system and operating on an Intel Core i9-12900K processor with 64 GB DDR-4 RAM and Nvidia RTX 3060 with 12 GB v-RAM. The hardware configuration for the low-power near-edge device was that of a Raspberry Pi 5 with 8 GB system memory. It has a Rasberry Pi OS installed on a Quad-Core ARM Cortex A76 processor clocked at 2.4 GHz and a Videocore-VII GPU clocked at 800 MHz. We also ran the model on an x64 Intel NUC running Ubuntu 18.04 LTS server with a Core-i5 7260U quad-core processor clocked at 2.2 GHz with Intel Iris®plus graphics 640 and 16 GB of system memory that acts as the on-premises device. The training of the individual models for each disease was performed at the far-edge server, and finally, the ensembling of these models was performed at the far-edge, near-edge, and on the on-premises machines for comparison. The Intel NUC provides a balanced perspective, considering both mobility and computational power. Meanwhile, the Raspberry Pi 5 represents a SWaP (Size, Weight, and Power) constrained mobile device, highlighting the practical aspects of deploying edge computing in resource-limited environments. We have also used a 5-fold cross-validation while training the model

using a batch size of 32. We begin with a learning rate of 0.001 and run the models until 50 epochs with the Adam optimizer using 80% of the data for training and the remaining 20% for validation. Some sample images from each class of diseases are shown in Fig.7. The distribution of images in the training and test sets is shown in Table 2.

TABLE 2. Detailed breakdown of the train and validation dataset.

Image Class	Dataset	
	Training Set	Validation Set
Normal	500	125
COVID-19	476	119
Viral Pneumonia	494	124
Total	1470	368

The images were further processed using the steps detailed in Fig 2. Furthermore, we have used image augmentation to avoid overfitting (overfitting occurs when a model learns the training data too well and cannot generalize to new data). Image augmentation artificially increases the size of the training dataset by creating new images from existing images. This helps the model learn more about the underlying patterns in the data and generalize new data better. The augmented images are shown in Fig. 4.

TABLE 3. Comparative analysis of the SqueezeNet and MobileNetV2 models for prediction on all 3 classes.

Model	SqueezeNet	MobileNetV2
Accuracy	0.9478	0.9753
Precision	0.9487	0.9752
Recall	0.9478	0.9753
F1 Score	0.9480	0.9752
AUC	0.9949	0.9982
Time	38m 22s	48m 25s

As a baseline, we trained the MobileNetV2 and SqueezeNet models on all three classes, and for the evaluation of the performance of our method, we used a train-cross-validation-test approach. The selection of models is highly dependent on these use cases. Specifically, the choice of MobileNetV2 and SqueezeNet is driven by their suitability for resource-constrained environments at the edge and on-premises. These models have fewer parameters, leading to faster training times and smaller model sizes, which are essential for efficient distribution and collaborative resource integration. Further, the fewer parameters in MobileNetV2 and SqueezeNet not only accelerate the training process but also ensure that the models can fit into the memory of devices with limited resources. This enables more flexibility for deployment in environments where computational efficiency and model size are critical factors. The training set was used to train the deep CNN model, the 5-fold cross-validation set was used to fine-tune the model's hyperparameters, and the test set was used to assess the proposed method's effectiveness. We trained the deep CNN model with a mini-batch size of 32. We used the Adam optimizer with weight decay, which is a technique that helps to prevent the model from overfitting. The initial learning rate was set to 0.001, and the maximum number of epochs to train the network was 50.

V. RESULTS AND DISCUSSION

Chest X-rays are the most common type of diagnosis tool for respiratory diseases, with over 2 billion in medical examinations per year. Nonetheless, we have a scarcity of radiology experts who can interpret these images. A significant benefit of developing machine learning models for respiratory disease diagnosis is the possibility of capturing and disseminating some aspects of expert knowledge from the labeling of the images used for training the models. These models are trained on large datasets that include annotations by experienced diagnosticians that could assist physicians all over the world by providing additional insights or second opinions during the diagnostic process. As shown in Fig. 8, SqueezeNet and MobileNet-V2 are the two models that have one of the least amount of trainable parameters and, hence, the simpler models for training them on the dataset. In Table 3, the results of the training process for all three classes using the CNN models identified earlier, namely MobileNetV2 and Squeezenet, are reported. These models provide excellent results in classification, as can be seen from the accuracy and loss curves in Fig. 9. However, they require a significant amount of resources in terms of computation and time for training, as described in Table 3, which does not provide real-time classification of diseases. Additionally, we need more complex models when we add more classes of diseases to our training dataset. Since no model is perfect for the classification of all classes of diseases, we would need to try out multiple models to get the best possible performance. To reduce the time and resources spent on training, we trained the models on individual diseases, and later, using the predictions from these smaller models, we inferred the predictions about the conditions of the patients using ensemble learning. These smaller models, trained for the classification of individual diseases, serve as the seed models for our ensemble models. They give us better results for the classification of a couple of classes of diseases, as shown in Fig. 10. This observation is corroborated in the complete performance metric shown in Table 4. We get almost perfect predictions for both of these diseases; however, the training time is still quite a significant bottleneck. The confusion matrix for each approach is shown in Fig. 12 to give the readers even more clarity about the classification performance of the two seed models. As we can observe in the confusion matrix, the two models perform commendably in classifying the diseases. Therefore, we have used these two models to make predictions using several ensemble learning approaches like average voting (Avg. Voting) and meta-learning approach using logistic (ML-LR) regression and decision trees (ML-DT), respectively.

As can be seen from Table 5, the average voting ensemble approach performs poorly when compared to the baseline models. This can be explained by the potential overfitting of the dataset, which results in subpar performance during validation. However, in contrast, meta-learning using the logistic regression (ML-LR) model results in performance that is in line with the SqueezeNet model for predicting the

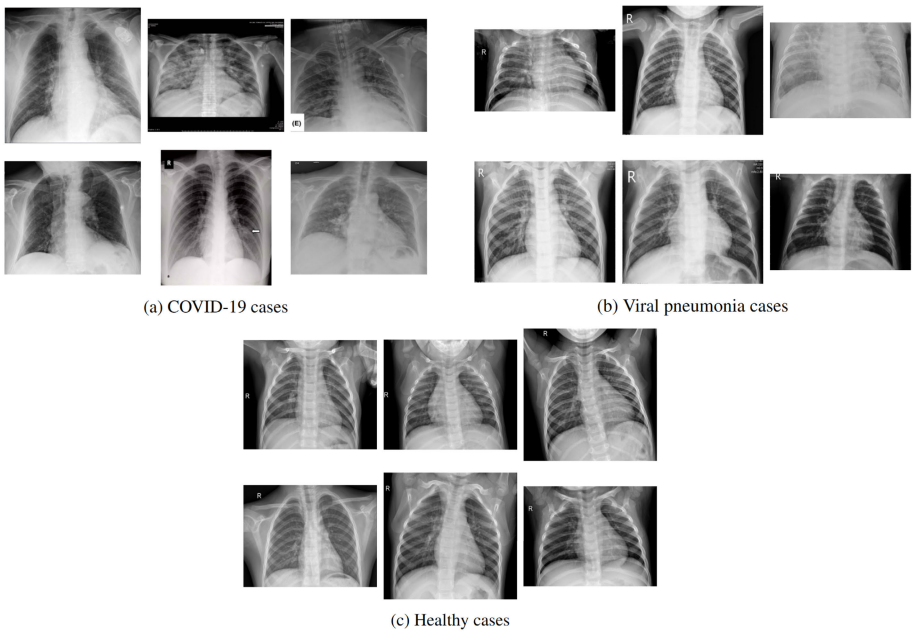


FIGURE 7. Compilation of three sets of chest X-ray images for different health issues. The first set (a) represents the X-rays of patients suffering from COVID-19. The second set (b) represents the X-rays of patients suffering from viral pneumonia. The last set (c) includes X-rays of healthy individuals without any sickness.

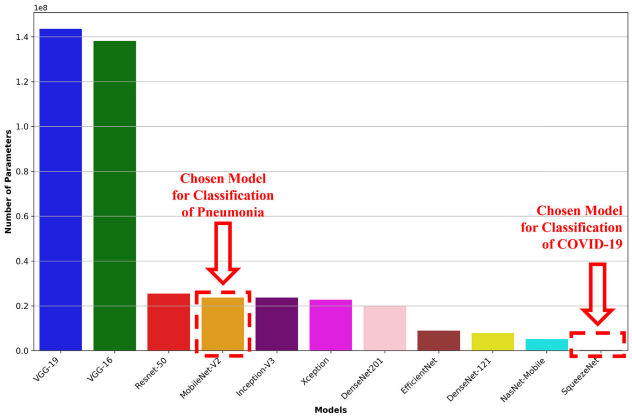


FIGURE 8. Number of trainable parameters in each model used for ensembling in the literature.

TABLE 4. Comparative analysis of the models for individual disease detection (Squeezenet for COVID-19 detection and Mobilenetv2 for Pneumonia detection).

Model	Squeezenet	Mobilenetv2
Accuracy	1.00	0.97
Precision	1.00	0.9758
Recall	1.00	0.9680
F1 Score	1.00	0.9719
AUC	1.0000	0.9932
Time	32m 28s	39m 17s

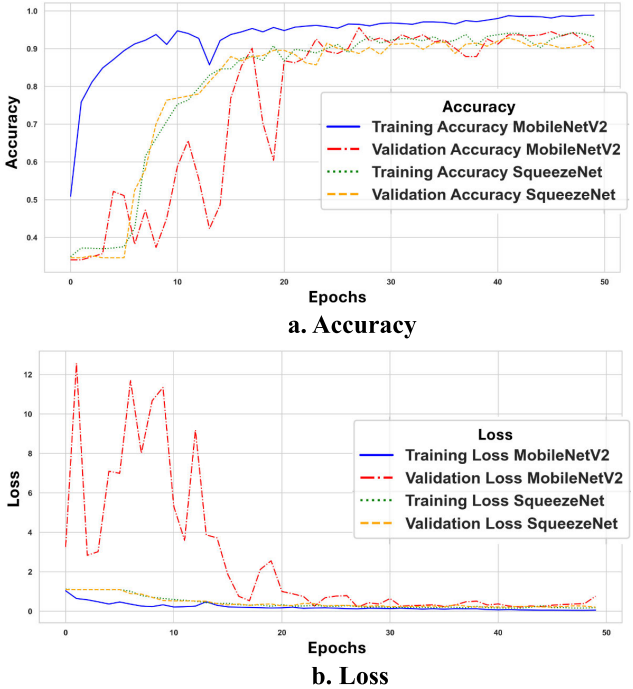


FIGURE 9. Performance evaluation for the baseline models classifying COVID-19 and pneumonia using the accuracy and loss plots during the training and validation phases.

output for the three classes of diseases. Moreover, the results were within 5% of the MobileNetV2 model. Nonetheless, when we use a decision tree (ML-DT) classifier, we get better and more explainable results, as shown in Fig. 11. This gives us a balance between the training and inference time as well as the critical metrics for evaluating the models. From the root node of the decision tree in Fig. 11, we can observe where the initial decision has been made after stacking the individual predictions column-wise. This aligns with our observations in Table 5. With a minimal overhead of (28 ms),

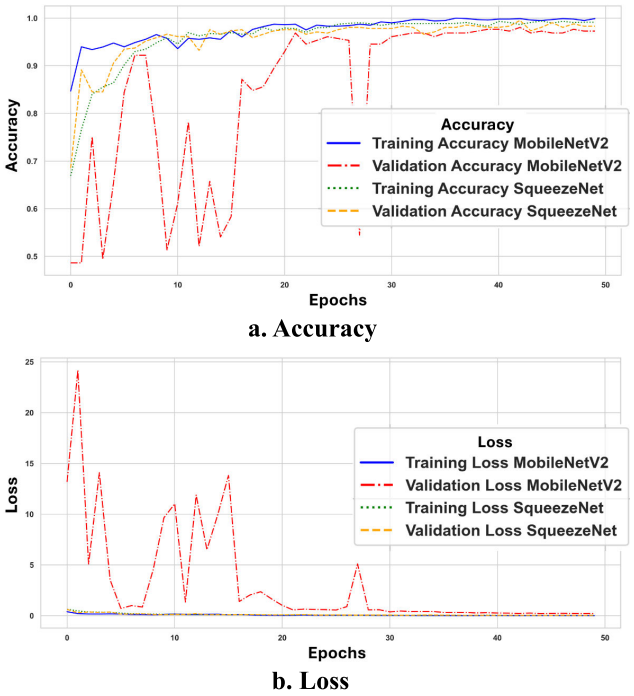


FIGURE 10. Performance evaluation of disease classification models for binary classification of the respiratory diseases using the accuracy and loss plots during training and validation phase using the seed models for ensemble learning.

we are reaching a perfect score in all of the crucial metrics. With accuracy, sensitivity, specificity, F1-score, and AUC of 97.80%, 97.06%, 98.48%, 96.51%, and 0.9739 respectively, we achieve commendable diagnostic performance. The ensemble model surpasses the sensitivity of only 60%-70% shown by the RT-PCR tests with a sensitivity of 96.48%. Additionally, we don't need manual intervention to set up the weight of the individual classification models, and we can make predictions based on the probabilities generated by the individual classification models. These smaller and lightweight models can effectively capture the details from the X-rays, thereby circumventing the high cost associated with manual labeling and hyperparameter tuning. Moreover, they can be trained independently over several geolocations and ensembled on-demand to preserve the privacy of the patients by on-site training of the models. This reduces the time required to anonymize the images for the training process and avoids the labeling errors introduced by the traditional NLP-based annotation systems used by the EHR software solutions. Therefore, it is our model of choice for deploying on the edge and on-premises servers.

TABLE 5. Comparative analysis of the three ensemble models.

Model	Avg. Voting	ML-LR	ML-DT
Accuracy	0.8379	0.9368	**0.9780**
Precision	0.8859	0.9440	**0.9706**
Recall	0.8379	0.9368	**0.9848**
F1 Score	0.8228	0.9369	**0.9651**
AUC	**0.9910**	0.9902	0.9739
Time	**7 ms**	151 ms	28 ms

In a resource-constrained environment such as hospitals that are being flooded with patients infected by novel diseases such as COVID-19, there is always a trade-off between the time required for an accurate prediction and the accuracy of prediction. Having trained our model on 1470 training images and testing the effectiveness of the model on 368 images, we are quite confident in its effectiveness. To further substantiate our claims, we have also compared our proposed approach with the previous studies described in the literature survey section, as shown in Table 6. We observe that our model performs well across all performance metrics, unlike the other models, where we see greater performance on one metric while compromising on others during evaluation. Moreover, the models used by the other studies have substantially more trainable parameters than the models we use. This can be confirmed by Fig. 8. Since these models perform binary classification, using a more complex model doesn't make sense, especially when we can achieve identical performance from smaller, lighter, and more efficient models. Furthermore, the additional complexity will result in longer training times, transmission delay, and larger memory needs during ensembling, which may be difficult to achieve on on-premises devices with less system memory.

As the final models were lightweight compared to the models used by other researchers, we explored the potential to run the ensembling process on low-power devices that we would typically find with the end users. We ran the experiments on lower-powered devices that usually act as portable mobile personal computers, near-edge, and far-edge devices, like a low-powered x86 PC, an ARM device, and a desktop equipped with a dedicated GPU, to confirm the assertion we made earlier. We measured the resource utilization on these devices as shown in Table 7. The lowest resource utilization and latency were observed on the commodity hardware desktop functioning as the far-edge server, which we used to train the CNN models earlier. This is representative of the performance we can expect from a far-edge device. Surprisingly, the Raspberry Pi 5, functioning as a near-edge device, performed much better than the Intel NUC in carrying out the task because of its higher clock speeds. We were able to run the meta-learning model and get modest performance in terms of computation time and

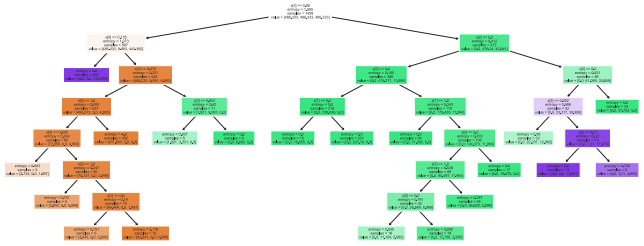
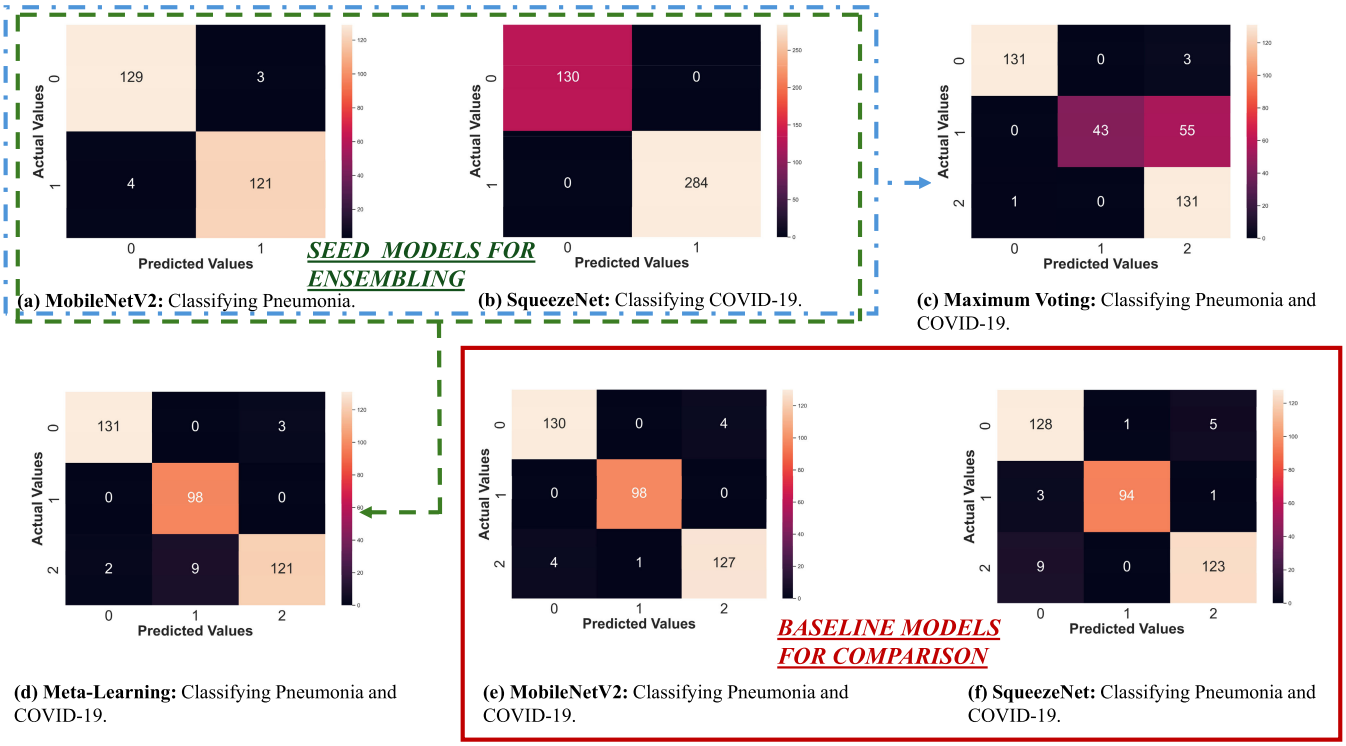


FIGURE 11. Decision tree of the ensembling model used for image classification. The tree is color-coded to represent different classes of images: orange nodes correspond to the first class, green nodes to the second class, and purple nodes to the third class. Each node in the tree represents a decision point, depicting the path taken depending on the outcome of the decision.



(a) MobileNetV2: Classifying Pneumonia.

(b) SqueezeNet: Classifying COVID-19.

(c) Maximum Voting: Classifying Pneumonia and COVID-19.

(d) Meta-Learning: Classifying Pneumonia and COVID-19.

(e) MobileNetV2: Classifying Pneumonia and COVID-19.

(f) SqueezeNet: Classifying Pneumonia and COVID-19.

FIGURE 12. Comparative performance of various machine learning models in classifying Pneumonia and COVID-19, as depicted by confusion matrices. Each matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering a comprehensive view of each model's predictive accuracy and error rate. The models are categorized into three groups: (a, b) Seed models (MobileNetV2 and SqueezeNet) trained on individual disease tasks to generate probability matrices for ensemble learning; (c) Maximum voting ensemble method combining predictions from seed models; (d) Meta-learning approach leveraging seed model outputs for joint disease classification. (e, f) Baseline models (MobileNetV2 and SqueezeNet) trained for simultaneous pneumonia and COVID-19 prediction. The meta-learning model (d) and standalone MobileNetV2 (e) emerge as the top-performing techniques, accurately classifying most cases across all disease classes.

TABLE 6. Comprehensive comparison of deep learning models for chest X-ray Diagnosis: sensitivity, specificity, accuracy, F1-Score, AUC and model parameters across various datasets.

Model	Sensitivity	Specificity	Accuracy	F1-Score	AUC	Model Parameters	Dataset	Year
CheXNeXt [13]	72.3%	67.4%	-	75%	0.862	69,682,060	NIH ChestX-ray14 set [59]	2018
Apostolopoulos and Mpesiana [17]	92.85%	**98.75%**	93.48%	-	-	167,519,029	COVID-19 image data collection [60] and CXR images [61]	2020
Majority voting classifier [36]	91.33%	86.21%	87.37%	**96.51%**	0.914	-	COVID-Chestxray set [60], Montgomery set [62], and NIH ChestX-ray14 set [59]	2020
CovXNet [37]	95.63%	91.25%	95.62%	91.21%	-	446,072,632	X-rays collected in Guangzhou Medical Center [61] and Sylhet Medical College COVID-19 dataset	2020
Iteratively Pruned Deep Ensembling [38]	**97.82%**	97.86%	**97.82%**	97.82%	**0.9969**	331,446,268	X-rays collected in Guangzhou Medical Center [61] and Sylhet Medical College COVID-19 dataset	2020
DeepCOVID-XR [14]	71%	92%	82%	-	0.90	145,512,785	NIH ChestX-ray14 set [59]	2021
RAIDER	97.06%	98.48%	97.80%	**96.51%**	0.9739	**24,272,882**	MIMIC-CXR and COVIDGR dataset	2024

other system metrics on a low-power ARM device, which justifies the deployment of our proposed architecture for practical applications. Our observations from Table 8 support these claims. We can process multiple samples in a second,

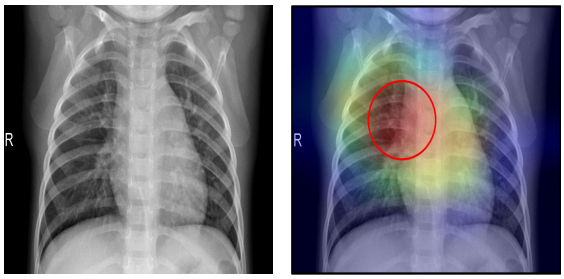


FIGURE 13. Comparative visualization pneumonia diagnosis- on the left side, we have an unprocessed X-ray of a lung belonging to a person who tested positive for Pneumonia. On the right, we have the highlighted GRAD-CAM heat map labeling the important parts of the lung that the model considers to be important for predicting Pneumonia.

TABLE 7. Resource utilization on various devices.

Machine	CPU	Memory		Time	
		RAM	v-RAM	Train	Test
Near-Edge	31%	2.8 GB	3.6 GB	23 sec	5 sec
Intel NUC	84%	3.4 GB	-	297 sec	49 sec
Raspberry Pi 5	98.10%	4.38 GB	-	153 sec	37 sec

TABLE 8. Throughput and network latency of different devices.

Machine	Throughput (samples/sec)		Latency per sample	
	Train	Test	Train	Test
Near-Edge	63.4348	72.8000	30.86 ms	28.77 ms
Intel NUC	4.9125	7.4286	204.3 ms	134.6 ms
Raspberry Pi 5	9.5359	9.8378	104.9 ms	101.6 ms

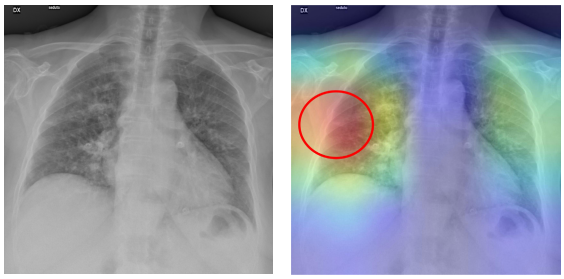


FIGURE 14. Comparative visualization COVID-19 diagnosis- on the left side, we have an unprocessed X-ray of a lung belonging to a person who tested positive for COVID-19. On the right, we have the highlighted GRAD-CAM heat map labeling the important parts of the lung that the model considers to be important for predicting COVID-19.

which assures us that these machines can keep up with the higher demands for diagnosis in real-life scenarios. Even if we assume that there's an average home internet connection of 300 Mbps at the diagnostic lab, which is a reasonable assumption to make in developing nations, we get a network latency of about 15.1 milliseconds for getting the results from the far-edge server, and an additional 15.76 milliseconds for processing. Even with this additional overhead for the network latency, we are getting the maximum throughput and minimum latency per sample for the near-edge server. The Intel NUC performs the worst, while the Raspberry Pi 5 sits in between. Thus, we can conclude that by using the near-edge and far-edge devices, we can get low latency predictions. However, if we are concerned about the privacy of the patient, then we can opt for the on-premises solution.

Finally, we also present the Grad-CAM images for different diseases in Fig. 14 and Fig. 13 for better interpretability

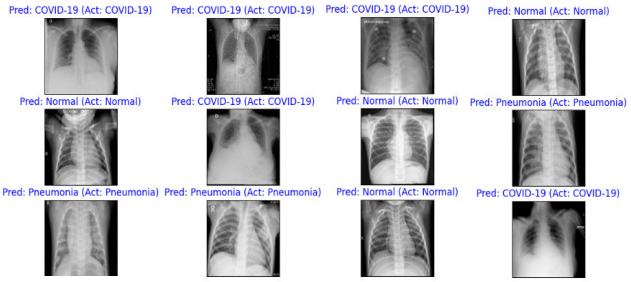


FIGURE 15. Post-training predictions with a series of X-ray images with the actual and predicted labels after training. The diversity of predictions across various conditions demonstrates the model's ability to differentiate between multiple diseases, underscoring its effectiveness in diagnosis.

of the models by the diagnosticians. The highlighted areas were identified as the most critical ones for making the classification. Fig. 15 shows several predictions made on the chest X-rays using this model with their respective truth labels and further justifies the balance RAIDER architecture has struck between training time and resource consumption.

VI. CONCLUSION

RAIDER has a scalable architecture with several collaborators working on improving the diagnosis of novel diseases. We have considered the deployment choices of an on-premises system for preserving privacy and an edge computing-enabled network. To provide comprehensive insights into the computational complexity, resource requirements, and practical deployment aspects, we evaluated the performance of training and testing across three distinct categories of devices, namely a desktop equipped with a discrete GPU that acts as the far-edge server, a portable and mobile PC (Intel NUC) that acts as the on-premises device, and an IoT device (Raspberry Pi 5) that acts as the near-edge server. As shown by the results of this study, we can use the ensembling model to quickly identify and diagnose new diseases and improve healthcare delivery during the onset of novel respiratory diseases that the chest radiographs can identify. We use the expert diagnosis of specialized radiology experts and share their expert diagnosis with other physicians across the globe. The models trained on these labeled images would be highly accurate at diagnosing respiratory diseases. We can deal with the shortage of radiology experts by providing a second opinion to the physicians to enhance the quality and accuracy of their diagnosis. RAIDER also provides high-precision and low-latency diagnosis with minimal system resource requirements. Using this framework, we are able to quickly and accurately diagnose the disease. Using meta-learning, we combined several models developed to diagnose individual diseases and precisely get the predictions for all of the classes of diseases present in our dataset without the need for manual weight assignment to the individual models, thereby reducing the bias in the final model. With the evolving needs of the healthcare industry, accurate and efficient diagnosis of diseases is the need of the hour. The

RADIER architecture provides an efficient and cost-effective solution for addressing these needs.

REFERENCES

[1] J. Wojtusiak, Y. Wang, V. Vakkalagadda, F. Alemi, and A. Roess, "Using Wi-Fi infrastructure to predict contacts during pandemics," in *PRoc. IEEE 9th Int. Conf. Healthcare Informat. (ICHI)* Aug. 2021, pp. 347–356.

[2] P. D. Yadav, G. N. Sapkal, R. Ella, R. R. Sahay, D. A. Nyayanit, D. Y. Patil, G. Deshpande, A. M. Shete, N. Gupta, V. K. Mohan, P. Abraham, S. Panda, and B. Bhargava, "Neutralization of beta and delta variant with sera of COVID-19 recovered cases and vaccinees of inactivated COVID-19 vaccine BBV152/Covaxin," *J. Travel Med.*, vol. 28, no. 7, Oct. 2021, Art. no. taab104.

[3] Y. Wang, P. Tran, and J. Wojtusiak, "From wearable device to Open-EMR: 5G edge centered telemedicine and decision support system," in *Proc. 15th Int. Joint Conf. Biomed. Eng. Syst. Technol. (BIOSTEC)—HEALTHINF*. Setúbal, Portugal: SciTePress, 2022, pp. 491–498, doi: 10.5220/0010837600003123.

[4] Y. Wang and T. Liao, "Data integrity and causation analysis for wearable devices in 5G," in *Proc. IEEE Int. Conf. E-Health Netw., Appl. Services (HealthCom)*, Oct. 2022, pp. 142–148.

[5] Y. Yi, P. N. P. Lagniton, S. Ye, E. Li, and R.-H. Xu, "COVID-19: What has been learned and to be learned about the novel coronavirus disease," *Int. J. Biol. Sci.*, vol. 16, no. 10, pp. 1753–1766, 2020.

[6] Y. Wang and J. Wojtusiak, *Active Learning Based User-Defined Chest X-ray Diagnosis System Leveraging 5G Infrastructure for COVID-19 Variant Detection*. Bethesda, ML, USA: AMIA, 2021.

[7] B. Udugama, P. Kadhiresan, H. N. Kozlowski, A. Malekjahani, M. Osborne, V. Y. C. Li, H. Chen, S. Mubareka, J. B. Gubbay, and W. C. W. Chan, "Diagnosing COVID-19: The disease and tools for detection," *ACS Nano*, vol. 14, no. 4, pp. 3822–3835, Apr. 2020.

[8] L. G. B. A. Quekel, A. G. H. Kessels, R. Goei, and J. M. A. van Engelshoven, "Detection of lung cancer on the chest radiograph: A study on observer performance," *Eur. J. Radiol.*, vol. 39, no. 2, pp. 111–116, Aug. 2001.

[9] F. Shi, L. Xia, F. Shan, B. Song, D. Wu, Y. Wei, H. Yuan, H. Jiang, Y. He, Y. Gao, H. Sui, and D. Shen, "Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification," *Phys. Med. Biol.*, vol. 66, no. 6, Mar. 2021, Art. no. 065031.

[10] 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.

[11] G. Bonito, V. Martinelli, F. Vullo, F. Basilico, E. Polito, A. Izzo, L. Corso, and P. Ricci, "Pictorial guide for variants of COVID-19: CT imaging and interpretation," *BJR| Open*, vol. 5, no. 1, Nov. 2023, Art. no. 20220011.

[12] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019, *arXiv:1901.07031*.

[13] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[14] R. M. Wehbe, J. Sheng, S. Dutta, S. Chai, A. Dravid, S. Barutcu, Y. Wu, D. R. Cantrell, N. Xiao, B. D. Allen, G. A. MacNealy, H. Savas, R. Agrawal, N. Parekh, and A. K. Katsaggelos, "DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. clinical data set," *Radiology*, vol. 299, no. 1, pp. E167–E176, Apr. 2021.

[15] I. Drozdov, B. Szubert, E. Reda, P. Makary, D. Forbes, S. L. Chang, A. Ezhil, S. Puttagunta, M. Hall, C. Carlin, and D. J. Lowe, "Development and prospective validation of COVID-19 chest X-ray screening model for patients attending emergency departments," *Sci. Rep.*, vol. 11, no. 1, p. 20384, Oct. 2021.

[16] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Sci. Rep.*, vol. 9, no. 1, p. 6381, Apr. 2019.

[17] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[18] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and F. Herrera, "COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 12, pp. 3595–3605, Dec. 2020.

[19] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[20] M. de la I. Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. González, and J. M. Salinas, "BIMCV COVID-19+: A large annotated dataset of RX and CT images from COVID-19 patients," 2020, *arXiv:2006.01174*.

[21] X. P. Burgos-Artizzu, "Computer-aided COVID-19 patient screening using chest images (X-ray and CT scans)," *medRxiv*, Jul. 2020.

[22] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, A. K. Denniston, H. Ashrafian, A. L. Beam, A.-W. Chan, G. S. Collins, A. D. J. Deeks, and M. K. ElZarrad, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," *Lancet Digit. Health*, vol. 2, no. 10, pp. e537–e548, Sep. 2020.

[23] T. Olatunji, L. Yao, B. Covington, A. Rhodes, and A. Upton, "Caveats in generating medical imaging labels from radiology reports," 2019, *arXiv:1905.02283*.

[24] E. Çallı, K. Murphy, E. T. Scholten, S. Schalekamp, and B. van Ginneken, "Explainable emphysema detection on chest radiographs with deep learning," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0267539.

[25] L. A. Smith, L. Oakden-Rayner, A. Bird, M. Zeng, M.-S. To, S. Mukherjee, and L. J. Palmer, "Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 5, no. 12, pp. e872–e881, Dec. 2023.

[26] H. Mahdyoon, R. Klein, W. Eyler, J. B. Lakier, S. C. Chakko, and M. Gheorghiade, "Radiographic pulmonary congestion in end-stage congestive heart failure," *Amer. J. Cardiol.*, vol. 63, no. 9, pp. 625–627, Mar. 1989.

[27] G. S. Francis, R. Cogswell, and T. Thenappan, "The heterogeneity of heart failure: Will enhanced phenotyping be necessary for future clinical trial success?" *J. Amer. College Cardiol.*, vol. 64, no. 17, pp. 1775–1776, 2014.

[28] S. Chakko, D. Woska, H. Martinez, E. D. Marchena, L. Futterman, K. M. Kessler, and R. J. Myerburg, "Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure: Conflicting results may lead to inappropriate care," *The Amer. J. Med.*, vol. 90, no. 1, pp. 353–359, 1991.

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 3462–3471, doi: 10.1109/CVPR.2017.369. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.369

[30] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 31–40, Jan. 2017.

[31] A. Ekbal and S. Saha, "A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14760–14772, Nov. 2011.

[32] J. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Using ensemble learners to improve classifier performance on tweet sentiment data," in *Proc. IEEE Int. Conf. Inf. reuse Integr.*, Aug. 2015, pp. 252–257.

[33] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, pp. 241–258, Apr. 2020.

[34] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Inf. Fusion*, vol. 37, pp. 132–156, Sep. 2017.

[35] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1249, Jul. 2018.

[36] T. B. Chandra, K. Verma, B. K. Singh, D. Jain, and S. S. Netam, "Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113909.

[37] 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.
[38] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays," *IEEE Access*, vol. 8, pp. 115041–115050, 2020.
[39] S. A. B. Ahmad, M. N. Taib, N. E. A. Khalid, R. Ahmad, and H. Taib, "The effect of sharp contrast-limited adaptive histogram equalization (SCLAHE) on intra-oral dental radiograph images," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Nov. 2010, pp. 400–405.
[40] E. Whaites and N. Drage, *Essentials of Dental Radiography and Radiology*. Amsterdam, The Netherlands: Elsevier, 2013.
[41] D. R. Ningsih, "Improving retinal image quality using the contrast stretching, histogram equalization, and CLAHE methods with median filters," *Int. J. Image, Graph. Signal Process.*, vol. 12, no. 2, pp. 30–41, Apr. 2020.
[42] P. P. Acharjya, S. Mukherjee, and D. Ghoshal, "Digital image segmentation using median filtering and morphological approach," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 1, pp. 552–557, 2014.
[43] H. A. Rahim, A. S. Ibrahim, W. M. D. W. Zaki, and A. Hussain, "Methods to enhance digital fundus image for diabetic retinopathy detection," in *Proc. IEEE 10th Int. Colloq. Signal Process. Appl.*, Mar. 2014, pp. 221–224.
[44] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. New York, NY, USA: Academic, 1994, pp. 474–485.
[45] J. Ma, X. Fan, S. X. Yang, X. Zhang, and X. Zhu, "Contrast limited adaptive histogram equalization-based fusion in YIQ and HSI color spaces for underwater image enhancement," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 7, 2018, Art. no. 1854018.
[46] K. B. Khan, A. A. Khaliq, A. Jalil, and M. Shahid, "A robust technique based on VLM and frangi filter for retinal vessel extraction and denoising," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192203.
[47] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Process.-Syst. Signal, Image, Video Technol.*, vol. 38, no. 1, pp. 35–44, Aug. 2004.
[48] H. Ju, D. Lee, S. Kang, and H. Yu, "Mitigating viewpoint sensitivity of self-supervised one-class classifiers," *Inf. Sci.*, vol. 611, pp. 225–242, 2022, doi: 10.1016/j.ins.2022.08.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025522009306
[49] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[50] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.
[51] L. Chaves, A. Bissoto, E. Valle, and S. Avila, "The performance of transferability metrics does not translate to medical tasks," in *Domain Adaptation and Representation Transfer: 5th MICCAI Workshop, DART 2023, Held in Conjunction With MICCAI 2023, Vancouver, BC, Canada, October 12, 2023, Proceedings*, vol. 14293, L. Koch, M. J. Cardoso, E. Ferrante, K. Kamnitsas, M. Islam, M. Jiang, N. Rieke, S. A. Tsaftaris, and D. Yang, Eds., Cham, Switzerland: Springer, 2024, pp. 105–114.
[52] E. T. Hassan, X. Chen, and D. Crandall, "Unsupervised domain adaptation using generative models and self-ensembling," 2018, *arXiv:1812.00479*.
[53] H. Kim, H. Kim, H. Moon, and H. Ahn, "A weight-adjusted voting algorithm for ensembles of classifiers," *J. Korean Stat. Soc.*, vol. 40, no. 4, pp. 437–449, Dec. 2011.
[54] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, and T. I. Ren, "META-DES: A dynamic ensemble selection framework using meta-learning," *Pattern Recognit.*, vol. 48, no. 5, pp. 1925–1935, May 2015.
[55] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward, "Improving predictions using ensemble Bayesian model averaging," *Political Anal.*, vol. 20, no. 3, pp. 271–291, 2012.
[56] C. Soares, P. B. Brazdil, and P. Kuba, "A meta-learning method to select the kernel width in support vector regression," *Mach. Learn.*, vol. 54, no. 3, pp. 195–209, Mar. 2004.
[57] S. Kuruvayil and S. Palaniswamy, "Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7271–7282, Oct. 2022.
[58] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, Feb. 2023.
[59] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.
[60] J. Paul Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," 2020, *arXiv:2003.11597*.
[61] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (OCT) and chest X-ray images for classification," *Mendeley Data*, vol. 2, no. 2, p. 651, 2018.
[62] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y.-X. Wang, P.-X. Lu, and C. J. McDonald, "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 233–245, Feb. 2014.

ISHAN ARYENDU (Graduate Student Member, IEEE) received the B.Tech. degree in computer science and engineering from Odisha University of Technology and Research, in 2018, and the M.S. degree from the Stevens Institute of Technology, Hoboken, USA, in 2023, where he is currently pursuing the Ph.D. degree with the School of Systems and Enterprises. His research interests include health informatics, collaborative communication in V2X systems, the age of information, and machine learning.

YING WANG (Member, IEEE) received the B.E. degree in information engineering from Beijing University of Posts and Telecommunications, the M.S. degree in electrical engineering from the University of Cincinnati, and the Ph.D. degree in electrical engineering from Virginia Polytechnic Institute and State University. She is currently an Associate Professor with the School of System and Enterprises, Stevens Institute of Technology. Her research interests include cybersecurity, wireless AI, edge computing, health informatics, and software engineering.