B2-Net: an artificial intelligence powered machine learning framework for the classification of pneumonia in chest x-ray images

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
7 January 2023

REVISED
14 February 2023

ACCEPTED FOR PUBLICATION
9 March 2023

PUBLISHED
3 April 2023

PAPER

B2-Net: an artificial intelligence powered machine learning framework for the classification of pneumonia in chest x-ray images

K M Abubeker* and **S Baskar**

Faculty of Engineering, Department of Electronics and Communication Engineering, Karpagam Academy of Higher Education, Coimbatore, India

* Author to whom any correspondence should be addressed.

E-mail: kmabubeker82@gmail.com

Keywords: artificial intelligence, depth wise convolution, data augmentation, ensemble learning, machine learning, pneumonia

Abstract

A chest x-ray radiograph is still the global standard for diagnosing pneumonia and helps distinguish between bacterial and viral pneumonia. Despite several studies, radiologists and physicians still have trouble correctly diagnosing and classifying pneumonia without false negatives. Modern mathematical modeling and artificial intelligence could help to reduce false-negative rates and improve diagnostic accuracy. This research aims to create a novel and efficient multiclass machine learning framework for analyzing and classifying chest x-ray images on a graphics processing unit (GPU). Researchers initially applied a geometric augmentation using a positional transformation function to the original dataset to enhance the sample size and aid future transfer learning. Models with the best accuracy, area under the receiver operating characteristics (AUROC), F1 score, precision, recall, and specificity are chosen from a pool of nine state-of-the-art neural network models. The best-performing models are then retrained using an ensemble technique using depth-wise convolutions, demonstrating significant improvements over the baseline models employed in this research. With a remarkable 97.69% accuracy, 100% recall, and 0.9977 AUROC scores, the proposed Bek-Bas network (B2-Net) model can differentiate between normal, bacterial, and viral pneumonia in chest x-ray images. A superior model is retrained using the chosen dense convolutional network-160, residual network-121, and visual geometry group network-16 ensemble models. The diagnostic accuracy of the x-ray classification unit is enhanced by the newly designed multiclass network, the B2-Net model. The developed GPU-based framework has been examined and tested to the highest clinical standards. After extensive clinical testing, the final B2-Net model is implemented on an NVIDIA Jetson Nano GPU computer. Healthcare facilities have confirmed the B2-Net is the most effective framework for identifying bacterial and viral pneumonia in chest x-rays.

1. Introduction

Pneumonia is a lung infection that progresses rapidly, and many microorganisms, such as bacteria, viruses, and fungi, can bring it on. As a potentially fatal infection, pneumonia must be taken very seriously. The Centers for Disease Control and Prevention reports that pneumonia is the top cause of mortality among children less than 5 years old across the globe. Meteorological factors, air pollution, and lifestyle choices such as being overweight or smoking have contributed to the increase in pneumonia cases. United Nations' Sustainable Development Goal-3.2 aims to reduce death rates for infants and young children; it can be possible by using advances in biosensor technology, machine learning (ML) techniques, and convolutional neural networks (CNNs). When a virus infects the respiratory system, it negatively influences oxygen and carbon dioxide gas exchange in the alveoli, leading to pneumonia. As a result, a broad range of viruses can cause viral pneumonia; nevertheless, the influenza virus in adults and the respiratory syncytial virus in

children account for the vast majority of cases. Some individuals can develop long-term respiratory difficulties and lung fibrosis from the infection, as seen with Covid'19 pneumonia and the Middle East respiratory syndrome [1]. Pneumonia is associated with many symptoms, the most frequent of which are a dry, hacking cough, trouble breathing, high fever, sweating, and shivering. Although these signs are common among people with viral illnesses and pneumonia, they are not unique. Even though those with diabetes, obesity or compromised immune systems have it far worse.

Despite the best efforts of physicians and radiologists, viral and bacterial pneumonia continue to have a similar clinical presentation, making diagnosis difficult. If a patient with pneumonia symptoms, a stethoscope examination of the lungs is the first frontline, followed by radiological procedures such as an x-ray and CT scan to confirm the diagnosis [2, 3]. In addition to helping distinguish between bacterial and viral pneumonia, radiography images continue to serve as the gold standard for diagnosing the disorder. The next step is detecting and categorizing lung infections using radiographic imaging using ML frameworks, artificial intelligence (AI), and CNNs. Using technological advances like biosensors, computer-aided diagnostics, and AI, pneumonia can be diagnosed, treated, and prevented with simple and efficient measures.

Without high-powered computer capabilities and effective AI and mathematical models, it is difficult for researchers and clinicians to distinguish between viral and bacterial pneumonia without false-negative results. The proposed research primarily used a data augmentation (DA) strategy to address the uneven distribution of data and the deficiency of training samples. Second, a transfer learning (TL)-based ensemble strategy called Bek-Bas network (B2-Net) is developed, and the feasibility of mathematical modeling on graphics processing unit (GPU) hardware is investigated. The algorithms are designed and deployed in NVIDIA's Jetson Nano GPU computer, and training, testing, and validation are done on various test image datasets. The devised portable pneumonia classification framework outperforms current approaches regarding clinical assessment criteria and false-negative rate. This research provides many significant contributions, including those listed below.

1. Using a graphics processing environment, researchers designed and implemented a unique, efficient multiclass network, B2-Net, for pneumonia classification in chest x-ray images.
2. DA, depth wise convolutions (DWCs), and the ensemble method handle the fundamental deep-learning (DL) problems of poor dataset quality and lack of availability.
3. Third, the NVIDIA Jetson Nano computer is used to deploy the developed B2-Net model and put it through its paces in several testbeds. Test results showed a 100% true positive rate (TPR) and the top performance across all assessment metrics.

Following this overview, the second chapter dives into the specifics of various ML and DL algorithms and the efficacy of GPU-based systems. The need and current developments in TL, augmentation, and ensemble approaches are explained in the next chapter, followed by the B2-Net architecture to enhance the model assessment criteria. The fifth chapter describes the various performance measures and deployment hardware and software, followed by the research's conclusion.

2. Related work

ML, DL, and TL have all been more popular in recent years as a method for addressing complex classification challenges in medical images to detect diseases, including covid'19, lung infection and liver cirrhosis [4].

The rising popularity of AI-based healthcare solutions over the last few years is a key driver that will boost the market for AI in medical imaging over the next years. One of the most remarkable developments of the last few years is the capacity to identify pneumonia from chest x-ray images automatically. Researchers have explored utilizing deep convolutional neural network (DCNN) models to identify pneumonia in several studies [5, 6]. Significant progress has been made in image classification, object detection, and disease detection from medical images due to the development of CNNs. The availability of big datasets and advanced DL techniques have enabled it to train various ML models. These models are effective in classifying chest x-rays, which has led to their rising popularity.

The TL approach has greatly simplified the procedure, which permits rapidly retraining a very DCNN with very few images. The residual network (ResNet), GoogLeNet, visual geometry group network (VGGNet), and MobileNet are the major classification methods for chest x-ray images; all use the TL technique in the DL framework. Pre-trained TL was used by Alqudah *et al* [7] and Oyelade *et al* [8] to make diagnoses of pneumonia. Lacruz and Vidarte [9] developed a computer-assisted diagnostic tool that uses DL models to detect pneumonia using chest x-rays with impressive accuracy. Mehmood *et al* [10] employed TL in the AlexNet framework to improve accuracy and efficiency when imaging histological lung and colon tissues.

Ensemble learning is a popular method that combines the predictions of multiple classifiers to provide a single prediction for a test sample. When the foundational classifiers' discriminative information is included, more accurate predictions can be made. Ensemble methods such as majority voting, average, and weighted average probability are widely employed in earlier research [11]. Researchers have used fine-tuning and TL strategies in GoogLeNet, ResNet, VGGNet, and dense convolutional network (DenseNet) models to select assessment measures such as accuracy, area under the receiver operating characteristics (AUROC) curve, F1-score, and precision. Using a weighted average ensemble approach and five-fold cross-validation, Kundu *et al* [12] found that DenseNet-121, GoogLeNet, and ResNet-18 performed the best on the Kaggle Radiological Society of North America (RSNA) chest x-ray dataset. Using a heatmap and AI, Liz *et al* [13] created a model for detecting pediatric pneumonia; the results demonstrate that the model improves on previous methods regarding area under the ROC curve (AUC) and TPR. Fraiwan *et al* [14] demonstrate a multi-class classification utilizing ensemble learning to automatically detect respiratory disorders using stethoscopic data.

In many fields, including medical imaging, small datasets continue to be a primary cause of subpar CNN performance and overfitting on training data. Numerous augmentation techniques have been proposed, such as random flips, rotations, and adding multiple types of noise [15, 16]. Sharma *et al* [17] and Stephen *et al* [18] were able to get an accuracy rate of 90.68% and 93.73% on the Kermany dataset by augmenting the data on CNN architectures for pneumonia classification. The training of CNNs has improved with DA using generative adversarial networks, which generate new data without using a specific augmentation strategy [19].

Multi-dimensional imaging, advanced scanning and imaging technologies, and the emergence of mathematical computations all contribute to a rise in processing time and cost of algorithm development. Computers with a central processing unit (CPU) have limited computing capability, but the GPU helps alleviate this problem. NVIDIA's Compute Unified Device Architecture is a widely used approach for programming GPUs that facilitates parallel processing. However, the most probable cause can be identified and proposed in this research to make up for the deficit in the medical industry's ability to identify diseases without false negative rate; advancements in GPU performance, ML algorithms, and AI-based systems are required.

Ukwuoma *et al* [20] built a hybrid explainable DL architecture for exact pneumonia disease detection utilizing chest x-ray images and convolutional ensemble networks and the transformer encoder method. They scored 99.21% for binary classification and 98.19% and 97.29% for multi-classification. Ayan *et al* [21] designed a computer-aided pneumonia detection system using seven prominent CNN models pre-trained on the ImageNet dataset. AUC, accuracy, and sensitivity were 95.21%, 90.71%, and 97.76%, respectively. Li *et al* [22] offer a DL-based technique with spatial attention superposition and multilayer feature fusion for chest x-ray (CXR) pneumonia detection using RSNA and AIRI CXR datasets. SAS-MFF-YOLO classified pneumonia and lesion on the AI Research Institute dataset with 88.1% accuracy, 98.2% recall, and 99% AP50. Hussain *et al* [23] suggest utilizing DL to identify COVID-19 CXRs from those of healthy persons and those with bacterial or viral pneumonia. Their method improved COVID-19 detection, triage, and monitoring.

The research literature that has already been found various issues and constraints, as stated earlier. Less varied datasets, higher false negative rate and false positive rate (FPR), and more expensive processing are some serious issues. To overcome these difficulties, it is essential to create a trustworthy classification model that is free of false negatives. This research presents a heuristic approach to classifying viral and bacterial pneumonia named B2-Net using DA, DWC, and majority voting ensemble methods.

3. Materials and methods

A DL model's performance depends on the quality of the underlying data collection and the neural network model. In most cases, the quality of the data used to train the network is the root cause of less-than-ideal outcomes, even when an excellent neural network model is used. Access to a large dataset is crucial since its size directly affects the performance of the DL model. One of the most typical difficulties in deploying ML in medical research is a lack of appropriate data. This is because obtaining such information takes time, ethical issues, and other legal formalities. DA and TL techniques are more appropriate technologies to fix these two problems. DA refers to a group of methods for synthesizing new data points from existing data or for synthesizing new training data from existing ones. This is achieved via DL models and other domain-specific approaches applied to the training data set to generate novel training instances.

Figure 1 illustrates the steps of the postulated B2-Net ensemble approach for the pneumonia image classification system developed on the Jetson Nano GPU computer. Four positional transformation functions (PTFs) are applied to the Kaggle x-ray image dataset as part of the DA procedure. Using TL and DA, the nine CNN models (AlexNet, DenseNet-121, DenseNet-160, MobileNet-V2, MobileNet-V3, ResNet-121,

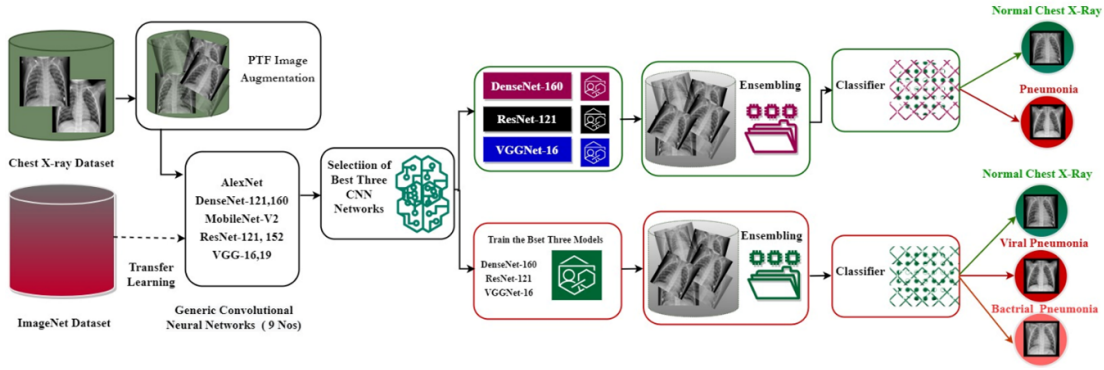

Figure 1. Graphical representation of the B2-Net model developed in AI powered ML framework for the classification of pneumonia in chest x-ray images.

ResNet-152, VGGNet-16, and VGGNet-19) are used to lessen the data imbalance and image deficiency. The three top models (DenseNet-160, ResNet-121, and VGGNet-16) are chosen based on performance measures such as accuracy, precision, recall, specificity, F1 score, and AUROC values.

The following section discusses a comprehensive investigation of various techniques and models employed in medical image analysis. Various types of CNN models are discussed in the first section, DA in the second section followed by TL and ensemble technique in the third and fourth sections respectively.

3.1. Fundamentals of CNN

As computing power increased exponentially and large datasets, GPUs, and DL algorithms became widely accessible, neural networks quickly replaced traditional approaches as the gold standard for virtually all computer vision-related tasks in research, academia, and industry. A CNN is used for both time series and image data to identify patterns and insights. It is excellent for tasks involving images, such as pattern recognition, object identification, and image classification. CNNs are structured similarly to the human brain, which has billions of neurons and is arranged so that the frontal lobe is the first to receive and analyze visual information. This configuration guarantees that the whole field of view is covered, solving the issue of feeding low-resolution image fragments into conventional neural networks. A CNN comprises three layers; convolutional, pooling, and fully connected (FC) layers. Within a convolution layer, a kernel or filter is swept over the image's receptive fields many rounds to identify the presence of an image feature. In this last layer, the image is quantified into a set of numbers that the CNN can use to decode and extract meaningful features. The pooling layer decreases the number of input parameters and causes some information loss by sweeping a kernel over the input image. In practice, adding a pooling layer to a CNN simplifies the network and makes it more effective. In a neural network, image categorization occurs at the FC layer using the characteristics acquired at earlier layers; this implies that every activation unit in the layer above directly connects to every input node below. The state-of-art models considered in this research are discussed in the following sections.

AlexNet is an eight-layer CNN network pre-trained on the ImageNet database [24]. The network has an image input size of 227 $\times$ 227, the first five convolutional layers, with max-pooling layers after the first, second, and fifth CNN levels. A dropout value of 0.5 is used in the first two FC layers. Except for the final one, the last three layers were FC layers that used a rectified linear units (ReLU) activation function. DenseNet was developed to optimize the information flow across layers. It has a different connectivity pattern, a direct connection from any layer to all subsequent layers. Each DenseNet composition layer performs pre-activation batch norm and ReLU, followed by 3 $\times$ 3 convolution with k-channel feature maps as output. As the connecting layers between two adjacent dense blocks, they utilize a 1 $\times$ 1 convolution followed by a 2 $\times$ 2 average pooling. A global average pooling (GAP) and a softmax classifier are subsequently appended at the last layer [25]. DenseNet 121 consists of two dense blocks, three translation layers, one classification layer, and five convolution and pooling layers

$$K^n = (K^0 + Kn - 1). \tag{1}$$

As shown in equation (1), the growth rate (K) determines how much new information is added to each successive layer, which controls the degree to which the ith layer can be generalized.

Inception-V3 image recognition model, achieving an accuracy of over 78.1% on the ImageNet database. This model's symmetric and asymmetric construction component are the same as the fundamental building blocks used in other cutting-edge algorithm designs. The model loss is calculated using the Softmax function,

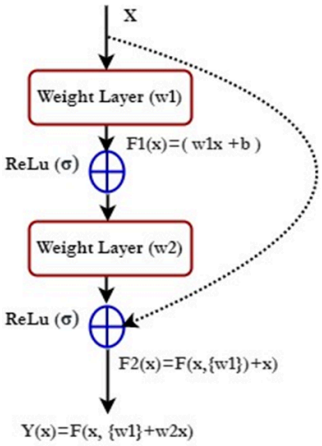

Figure 2. Skip connection used in ResNet architecture.

and batch normalization is heavily used. The MobileNet model employed depthwise separable convolutions (DSCs) and was developed by Google and made publicly available. Compared to a network using standard convolutions of the same depth, the number of parameters in this method is drastically reduced. Combining the benefits of two techniques, DWC, and pointwise convolution (PWC), a DSC improves CNN's ability to predict images, yields faster reaction times, and positions them to compete with mobile systems. In addition to these improvements, MobileNet-V2, and MobileNet-V3, provide new mobile network possibilities.

ResNet is a popular and very successful DL model of residual blocks with skip connections to deal with the vanishing gradient issue. When activations in one layer are linked to those in subsequent layers through a skip connection, the intermediate layers are ignored to create a residual block. Figure 2 below illustrates the idea of skip connection.

A layer's weights *(W)* are multiplied by the input (x) when no skip connection exists, and a bias (b) term is also included. Consequently, the function of output $(y(x))$ is determined by,

$$F1(x) = (w1x + b) \tag{2}$$

$$F2(x) = (w1x + w2x + b) \tag{3}$$

$$F2(x) = [(x, \{w1\}) + b)]. \tag{4}$$

When a new skip connection technique is applied the output $y(x)$ with ReLU activation function σ is given as,

$$Yx = Fx, W1 + W2x. \tag{5}$$

Here,

$$F = W2\sigma W1x. \tag{6}$$

The input and output dimensions cannot be the same when using a convolutional or pooling layer. Two methods exist for dealing with this issue: use zero padding with the skip connection to increase its dimensions, and include a 1×1 convolutional layer into the input to match the dimensions. Among the several ResNet designs, such as ResNet-50, 101, and 152, the ResNet-152 architecture is employed in this research for chest x-ray classification.

VGGNet uses larger kernel-sized filters than AlexNet and a series of 3×3 filters. Based on the number of CNN layers, there are two different VGG networks: VGGNet-16 and VGGNet-19 include 13 and 16 CNN layers and 3 FC layers, respectively. A 224×244 RGB image is supplied as VGG's input with a 3×3 or 1×1 filter and a fixed convolution step. Five max-pooling layers, out of a total of 16, are responsible for the spatial pooling; these levels follow certain conv. layers but not others. The max pooling and soft-max layer is the last one, and the hidden layers have a ReLU function.

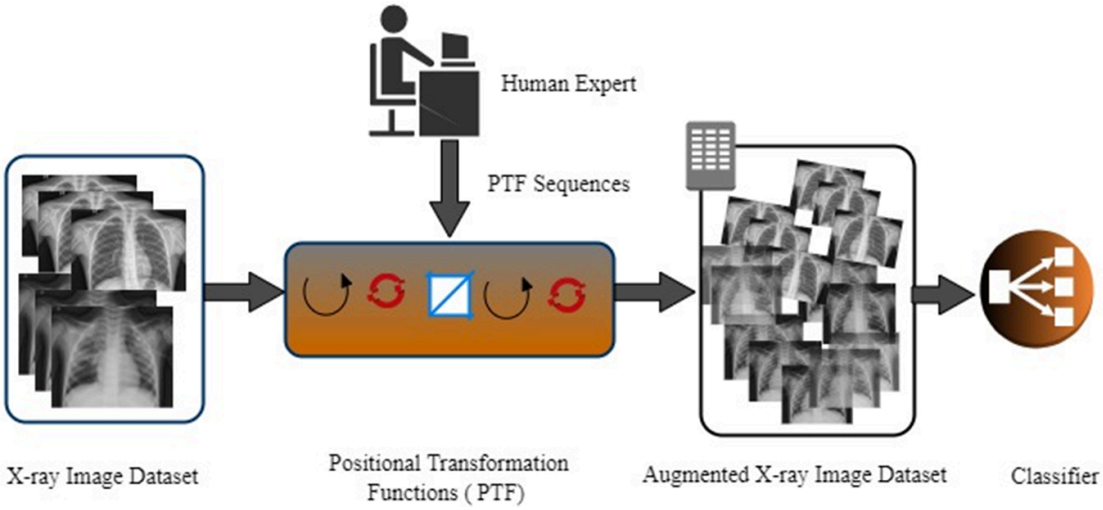

Figure 3. A heuristic data augmentation technique used in this B2-Net approach.

3.2. Fundamentals of DA

Today's deep neural networks (DNNs) and other sophisticated ML models can have billions of parameters and require enormous labeled training datasets that are often unavailable. As a result, DA, the synthetically growing labeled training datasets, has emerged as a crucial tool for overcoming this data scarcity issue. Every modern model for image classification today uses DA as a secret sauce. It is also common in other modalities like medical image classification, real-time object classification, agricultural areas, and natural language processing (NLP). DA aims to improve the generalizability of an overfitted data model. An overfitted data model ensues when a small subset of the data points overly constrains a function. Through the generation of extra training data and the model's exposure to multiple versions of already-existing data, DA is regularization and aids in managing data overfitting. It allows us to make the dataset more comprehensive and diverse without acquiring more data. DA aids in data cleansing and makes ML models more resilient by generating model variants. Curating datasets is impractical for medical imaging applications due to the high cost and time commitment involved in getting many annotated samples from specialists. Compared to the predicted differences of the identical x-ray images, the network trained with augmentation must be more robust and accurate. Figure 3 illustrates a unique heuristic approach to DA that has developed as an outcome of this research.

Figure 3 depicts the results of applying many basic PTFs on the original dataset, including rotations, v-shifts, h-shifts, and crops. DA techniques customized by humans can increase model performance. Heuristic DA strategies often depend on the construction of transformation functions. Positional transformation is used to improve the quality of predictions and training efficiency. The vast amount of publicly available image resources makes it feasible to include more x-ray images into the dataset used for training, testing and validating the system. However, practical applications of such heuristic procedures might result in significant variations in final model performance and can fail to provide enhancements required for state-of-the-art models.

3.3. Fundamentals of TL

There is a requirement for large amounts of data in the training phase of CNN models to increase their generalizability. However, gathering sufficient data is sometimes impossible for many computer vision issues. Collecting and categorizing information on medical issues takes a lot of time and effort. Several solutions to this issue have explored, which is encouraging. One of these techniques is DA, which boosts the model's accuracy, generalizability, and resistance to overfitting. The second method, known as TL, is linked to multi-task learning and concept drift concerns. On the other hand, TL is often utilized in DL owing to the large amounts of time and energy required to build DL models or the complexity of the datasets used to train these models. The process of improving a model's capacity to learn a new task from existing knowledge obtained while doing a related job is referred to as TL.

Figure 4 illustrates a visual representation of the overall structure of the TL technique. There are two ways to apply prior knowledge to the new endeavor. In the first, the internal weights of a previously trained model are frozen, and the model is used in its feature extractor role. A classifier is trained on this frozen architecture to complete the knowledge transfer and preserve as much of the original domain's information as feasible.

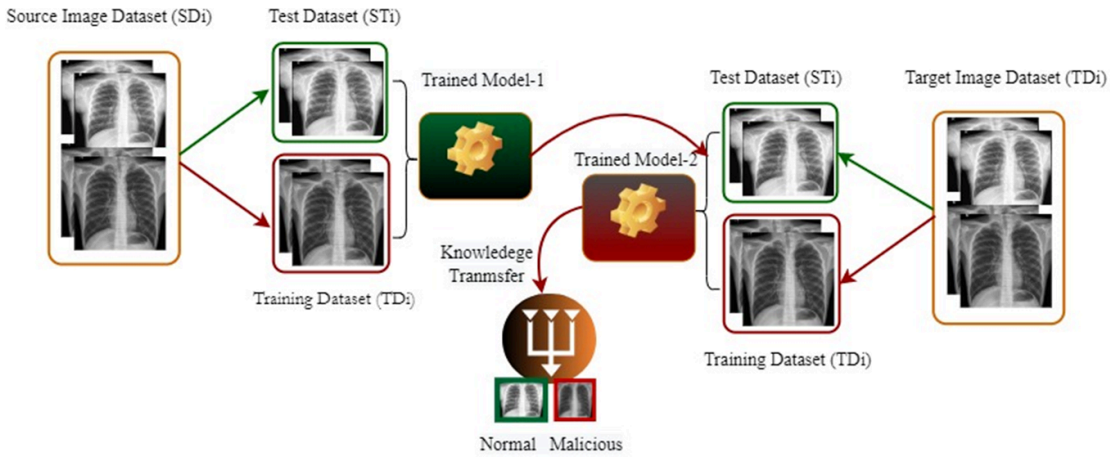

Figure 4. Workflow of transfer learning technique adopted in B2-Net model for the nine convolutional neural network models.

The second strategy focuses on optimizing the whole architecture for the new job. Instead of a completely arbitrary beginning point, the trained model's weights are employed here. The network can tailor its whole structure to the new domain and quickly obtain training convergence.

Using a feature extractor, fine-tuning, or pre-trained models are just a few examples of how TL can be implemented in DNNs. When training a network, it is important to consider whether the data has similarities with ImageNet data before settling on a specific method. Elements like edges, corners, and colors, which are not reliant on the data, appear in the first levels of a CNN, whereas features like textures, which are more specific to the data, appear in the later layers. Thus, various issues and data sets can benefit from using different first-layer kernels. Some initial convolutional layers are preserved in the fine-tuning process by not changing their weights while the model is trained.

In the first phase of the work, TL and fine-tuning processes were used to train nine well-known CNN models to assess whether or not a CXR image portrayed pneumonia or a normal one. Many hyperparameters like epoch size, dropout rate, and learning rate are adjusted when training a CNN model using the fine-tuning technique. Following the completion of the tests, the optimum values for the model's hyper-parameters and the weights that correspond to them were retained. After running each of the six CNN models through a series of tests, selected the three that performed the best for the data ensemble technique.

4. B2-Net model

The primary motivation for this research is to develop a portable gadget powered by a GPU capable of efficiently analyzing chest x-ray images and classifying them as either viral or bacterial pneumonia using a combination of different DCNNs. DWCs, squeeze-excitation, TL, DA, and weighted majority vote ensemble classification are the cornerstones of the approach used in this research.

Rather than relying on a single ML model, ensemble models pool the conclusions drawn by numerous models to boost global performance metrics like predictability and accuracy. Boosting, bagging, and stacking are the most well-known ensemble techniques; they are beneficial for regression and classification since they help decrease bias and variance and increase the accuracy of models. The minimization of overfitting, a problem with many predictive models, is achieved by decreasing the variance and improving accuracy. Boosting is an ensemble method that dramatically enhances the predictability of models by combining numerous weak base learners into a single strong learner. The stacking or stacked generalization approach functions by enabling a training algorithm to ensemble the predictions of many comparable learning algorithms. Regression, DL, and classifications are just a few areas where stacking has been effectively used.

The research experiments were conducted on Kaggle RSNA datasets since they are publicly accessible and can be accessed without cost. There are 6020 CXR images, and they are evenly distributed among three sets for training, testing, and validation. JPEG versions of pneumonia and normal images are stored in separate subfolders under each main heading. The training dataset has a total of 5216 images, with 3875 images categorized as pneumonia and 1341 as normal. There are 624 images in the test dataset, 390 of which are categorized as pneumonia and 234 as normal. Dataset distribution information utilized in this work is given in table 1.

Table 1. Details of Kaggle RSNA dataset [26] used in this research.

X-ray image type	Normal images	Bacterial pneumonia	Viral pneumonia	Total
Training	1341	1678	2197	5216
Testing	234	184	206	624
Validation	76	48	56	180
Total	1651	1910	2459	6020

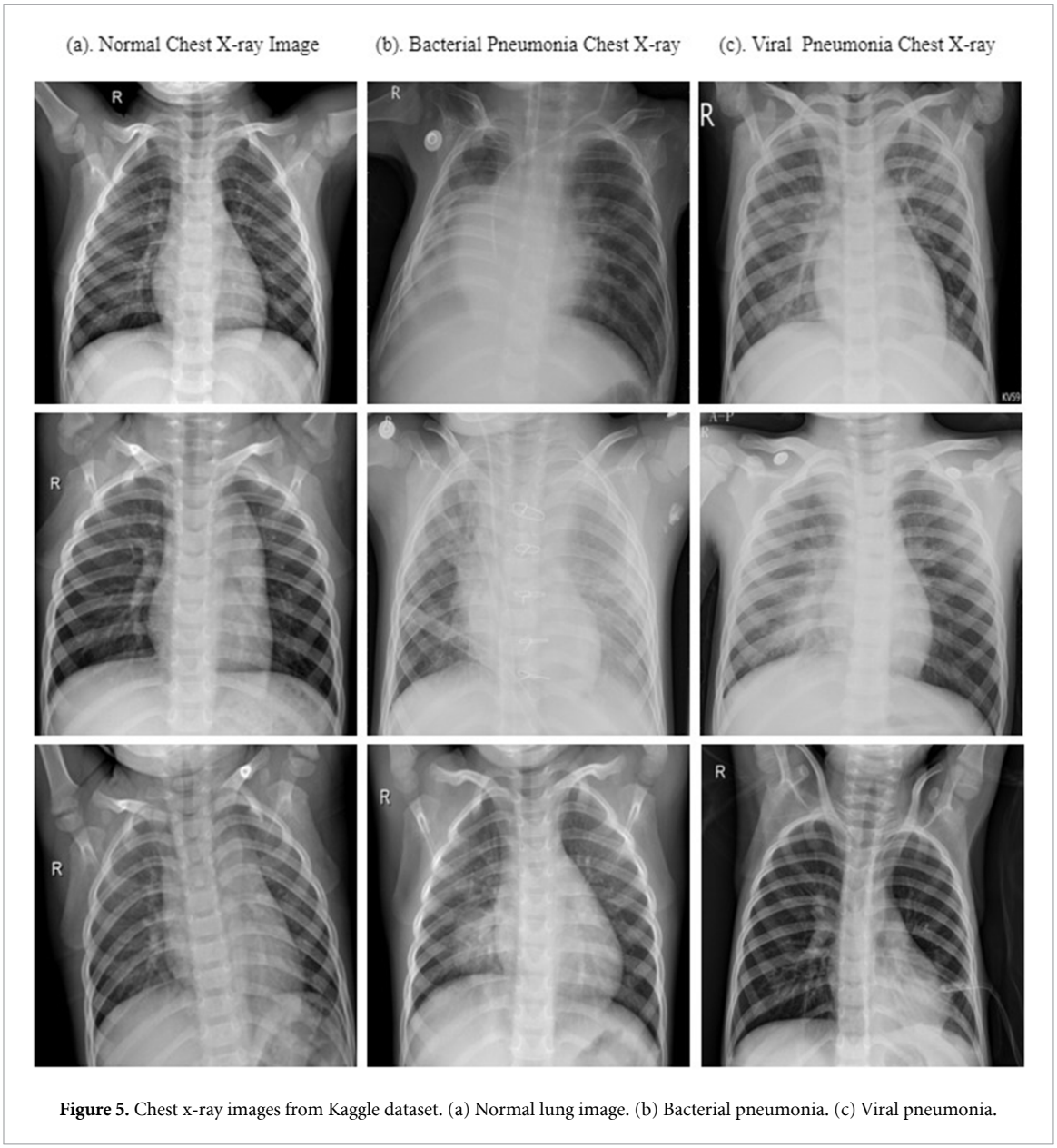

Figure 5. Chest x-ray images from Kaggle dataset. (a) Normal lung image. (b) Bacterial pneumonia. (c) Viral pneumonia.

Before being fed into the models, the chest x-ray images are scaled to 224 × 224 so that they can fit the input size of the architectures selected for the models. After further subdividing the x-ray images based on the infection source, researchers were left with three distinct classes (training, testing, and validation) in addition to those described before. The primary motivation for creating a separate validation set inside the dataset is to reduce the likelihood of the model overfitting. The training set must include various inputs to ensure the model is prepared to forecast any unknown data sample and the same epoch of training data is continuously fed into the CNN design. When training a model, its efficacy can be verified by comparing it to an independent validation set. The validation data is used to fine-tune the model's hyperparameters and other features. At the end of each epoch, the model is evaluated on the validation set and compared to its previous state, which is learned from the training set. Figure 5 shows the images taken from the Kaggle dataset used in this research.

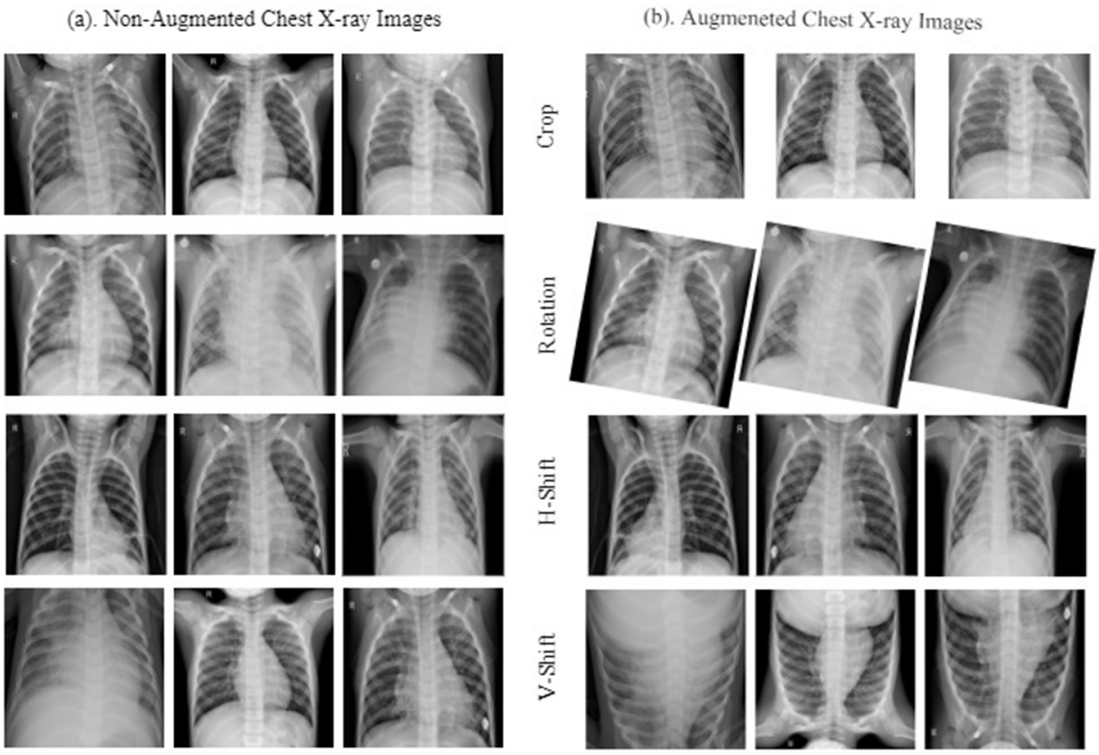

Figure 6. Images used in this research with and without augmentation in chest x-ray images.

Table 2. Details of augmented chest x-ray dataset used in this research.

X-ray image type	Normal images	Bacterial pneumonia	Viral pneumonia	Total
Training	2145	1678	2197	6020
Testing	234	185	238	657
Validation	192	156	190	538
Total	2571	2019	2625	7215

The model accuracy, generalization ability, and avoiding overfitting are all enhanced through augmented data. The DA technique is utilized in this research both before and during the training phase, as shown in figure 6.

Compared to the number of images labeled with pneumonia in table 1, the number of normal-labeled images was 6020. To bridge this gap and prevent overfitting, 1195 new augmented images were attained by randomly applying v-shifts, h-shifts, rotation by 10°, and finally, crop to the training dataset. Before and throughout the training phase of this investigation, the DA technique is used. The CXR images are processed using image analysis software to narrow the discrepancy. After augmentation applied to the training and testing dataset, resulting in 2571 normal 2019 bacterial and 2625 viral pneumonia images. Upon inspection of the dataset after augmentation, it was discovered that there are 6020 total images, 1195 pneumonia-labeled images than normal-labeled images.

Table 2 summarizes the x-ray image employed in the different steps of this B2-Net framework after DA. After examining the ratios of the datasets in tables 2 and 3, the ratio of datasets used in the training, testing, and validation datasets is formatted as an 80:10:10 distribution. The model keeps learning data characteristics by repeatedly being given the same training data in each epoch. While training, the model's performance is checked against the validation set, which is a different data collection. The validation results allow us to fine-tune the model's hyperparameters and other settings. After each epoch, the model is evaluated on the validation set and compared to its previous state, which is learned from the training set. Creating a separate validation set from the dataset is necessary to avoid overfitting the model. After training is complete, the model is put to the test using a second dataset called the test set, which yields objective metrics for the model's ultimate performance in terms of accuracy, precision, and so on.

DNNs employ several TL techniques, including feature extraction and fine-tuning approaches. The task being performed and the characteristics of the ImageNet dataset determine which TL network is used. Since

Table 3. Performance comparison of nine models in normal vs bacterial pneumonia.

SI. no.	Model	Accuracy	Precision	Recall	Specificity	F1-score	AUROC
1	AlexNet	89.41	82.30	89.58	89.31	85.79	0.9124
2	DenseNet-121	90.15	82.94	91.15	89.60	86.85	0.9481
3	DenseNet-160	97.40	95.88	96.88	97.69	96.37	0.9732
4	MobileNet-V2	90.89	83.89	92.19	90.17	87.84	0.9501
5	MobileNet-V3	91.82	85.24	93.23	91.04	89.05	0.9594
6	ResNet-121	96.10	93.40	95.83	96.24	94.60	0.9816
7	ResNet-152	94.98	91.88	94.27	95.38	93.06	0.9700
8	VGGNet-16	97.03	94.90	96.88	97.11	95.88	0.9916
9	VGGNet-19	94.80	91.84	93.75	95.38	92.78	0.9665

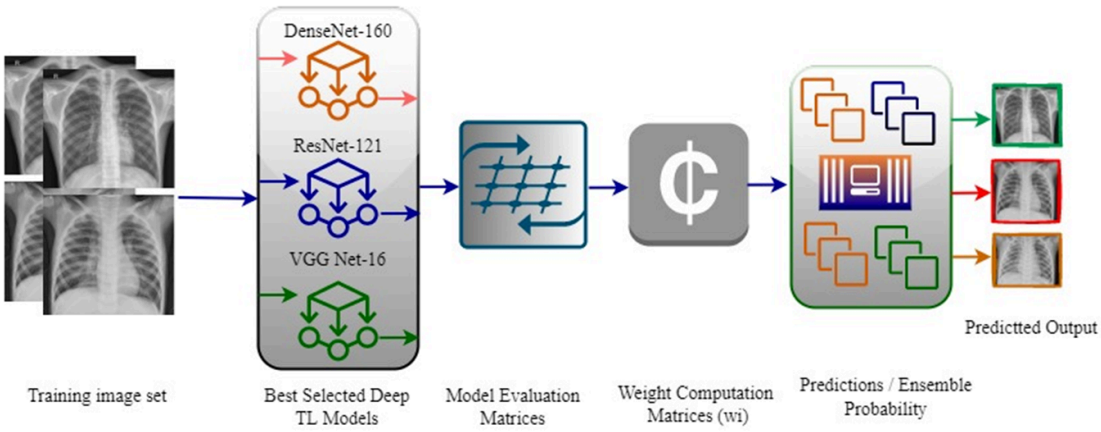

Figure 7. B2-Net-ensemble technique applied to the selected CNN models (DenseNet-160, ResNet-121, and VGG-16).

general characteristics manifest in the first-layer kernels, they are adaptable to various datasets and challenges. Some of the original CNN in the model are kept static throughout the fine-tuning phase, so their weights do not change while the model is trained. A two-pronged approach to TL is employed here. Researchers use TL and fine-tuning methodologies in the first phase to train nine popular CNN models to differentiate between pneumonia and normal chest x-ray images. To start training, images are downsampled to 224 × 224, the convolutional layers' weights are drawn from the pre-trained ImageNet weights, and a new classifier is learned from scratch. To get a superior model from the nine models mentioned earlier, researchers fine-tune their hyperparameters, including the number of frozen CNN layers, FC layers, dropouts, the optimization technique, the learning rate, and the epoch size. The best model for pneumonia classification using the extended dataset is determined to be a combination of DenseNet-160, ResNet-121, and VGGNet-16. To prevent overfitting, a dropout of 0.5% is applied to the layers in every CNN model before the FC layers. In addition, the issue of internal covariate shift is addressed by using batch normalization. Overfitting is kept at bay by decreasing the weights of the FC layer by a factor of 0.0001 using the L2 regularization technique.

In figure 7, the schematic depicting the implementation of the ensemble method used in this research. Neural network models are trained using TL and augmentation techniques on a total of nine models in this work, with the three best models being employed for the final model creation. During training, some models performed better in identifying normal x-ray images, while others performed better in identifying viral and bacterial pneumonia samples. ML classifiers that are either conceptually related or completely distinct are combined by majority vote. To improve classification accuracy, using a majority voting ensemble approach is recommended. Equation (7) shows that a majority vote across all classifiers can be employed to forecast the class label Y

$$Y = \text{mode}[C1(x), C2(x), \ldots, Cm(x)]. \tag{7}$$

DenseNet-121 ResNet-152 and VGGNet-19 to categorize a training sample as follows;

$$Y = [\text{mode}[\text{ResNet121} = 0, \text{DenseNet160} = 1, \text{VGGNet16} = 1]] = 1. \tag{8}$$

In addition, a weighted majority vote associated weight Wj with a classifier Cj and characteristic function XA, as given in equation (9)

$$Y = \text{argmax} \sum_{i=1}^{m} WjXA(Cj(x) = i) \tag{9}$$

where $XA = [Cj(x) = i \in A]$ then,

$$Y = \text{argmax}[W1 * i0 + W2 * i0 + W3 * i1] = 1 \tag{10}$$

$$Y = \text{mode}\,[\text{ResNet121} = 0\,|W1,\ \text{DenseNet160} = 1|\,W2,\ \text{VGGNet16} = 1|W3]. \tag{11}$$

However, the most crucial component in guaranteeing the improved performance of the ensemble is the weights (Wi) assigned to each classifier. Most methods in the literature use experimental data to determine the weight value, while many other researchers take into account the classification accuracy when allocating weights to the base learners. This practice becomes even less reliable when datasets are skewed toward one class or another. An innovative and effective weight allocation strategy to mitigate this issue is shown in figure 7; it assigns weights to the three foundational CNN models (DenseNet-160, ResNet-121, and VGGNet-16) based on their performance on six evaluation metrics.

Ensemble techniques are used to increase model performance and minimize error rate from using a single classifier. Weighted voting provides each classifier a specific degree of power based on predefined criteria, then tallies the votes. Here the weight Wi assignment of each classifier based on its training set accuracy. Equation (12) is used to calculate weight each criterion

$$Wi = Ai \Big/ \sum_{i=1}^{n} Ai. \tag{12}$$

In equation (12), Wi and Ai is the weight and accuracy of the classifier respectively, $\sum_{i=1}^{n} Ai$ is the average weight of the model accuracy. In general, model weight Wm for the six evaluation parameters ($P1$, $P2$,… $P6$) are calculated by the equation given below

$$MW = Pi \Big/ \sum_{j=1}^{6} Pi. \tag{13}$$

Similarly, the average weight of all the selected models are given by the equation (13) below

$$W_{\text{avg}} = MWi \Big/ \sum_{i=1}^{n} MWi. \tag{14}$$

As an approach of diagnosing pneumonia from chest x-rays, an innovative and efficient CNN architecture called B2-Net has been conceived and deployed on a GPU platform. Figure 7 depicts the basic layout of the B2-Net model, which consists of six convolutional blocks and a single output layer. It has one GAP layer, two softmax layers, four squeeze and excitation blocks, and ten DWC blocks. Since depth wise separable convolution (DWSC) minimizes the memory and compute bandwidth needs for convolution in neural networks, it is often utilized for neural networks that operate on edge devices. As opposed to the traditional convolution, DWC performs spatial convolution on each input channel separately. The total number of multiplications performed using N filters with sizes $Dk \times Dk \times C$ in a DWC is

$$C \times Dk2 \times Dp2. \tag{15}$$

In PWC, the kernel size is denoted by Dp, but in DWC, it is denoted by Dk.

The B2-Net model is plotted in figure 8; here B2-Net divides the input image ($224 \times 224 \times 3$) into three distinct hues (R, G, B). Channels of the input image are independently processed using a kernel divided into three parts ($k = 3$) during the input phase. There are five filter size restrictions, beginning with 32 and progressing upwards through 64, 128, 512, and 1024. The output tensor of the input image channel is constructed by generating the convolved feature map for each color channel of the input image and stacking them. In this B2-Net model, the three channels are integrated to form a single tensor, and their output is the combined sensor that the DWC framework will use.

For the network to alter the weight of each feature map, researchers need to add parameters to the channel of each convolutional block using squeezing and excitation. The size of the input block into the convolution layer is given by the (B, C, H, W). The height and breadth of the feature map are H and W,

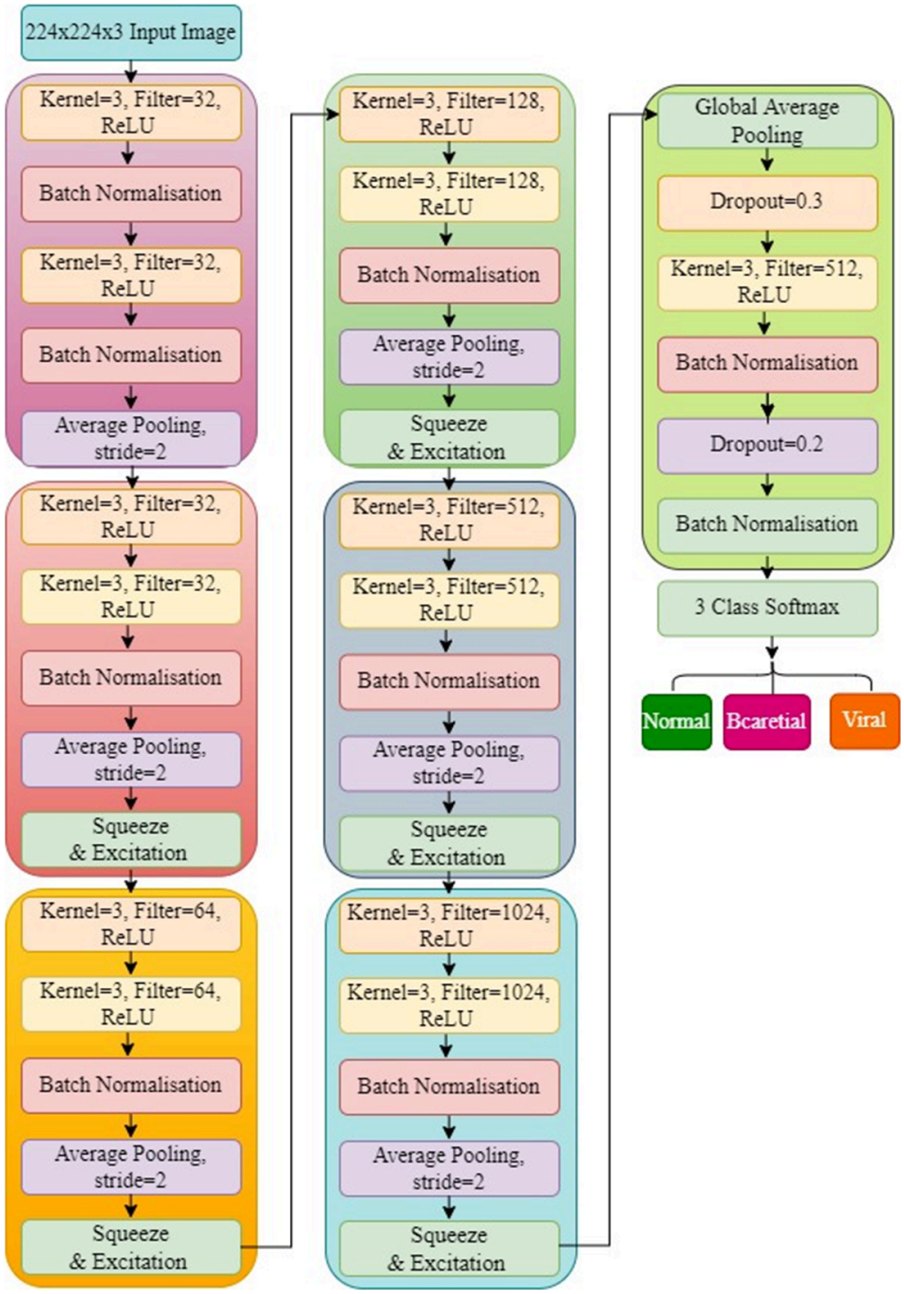

Figure 8. Architecture of B2-Net model developed and deployed in GPU platform for multiclass pneumonia classification from chest x-ray images.

where B and C stand for the batch size and channels, respectively. Squeeze models use GAP to average all the feature map's pixels into a single value. As a result, if the input tensor is of the form (CHW), the output tensor generated after applying the GAP function will be of shape $(1 \times 1 \times C)$. The squeeze-excitation block's effect on the CNN layer is shown in figure 9.

After the regular convolution operation with 3×3 filters, DWC used one batch normalization and one average pooling layer. As a result of the DWCs, the computational cost has been reduced, and researchers have been able to employ the squeeze and excitation block with a residual connection to extract relevant feature channels while ignoring irrelevant data and to conclude the second block with a GAP layer. The B2-Net was trained from scratch with 100 epochs at a 0.0001 learning rate and evenly distributed datasets. The depth-wise convolutional layer consists of a 3×3 DWC with a batch norm and ReLU, followed by a 1×1 PWC, followed by a batch normalization and ReLU. A 1×1 convolution operation is applied to the C channels whenever a pointwise action is performed. As a result, the size of the filter used for this operation will be $1 \times 1 \times C$.

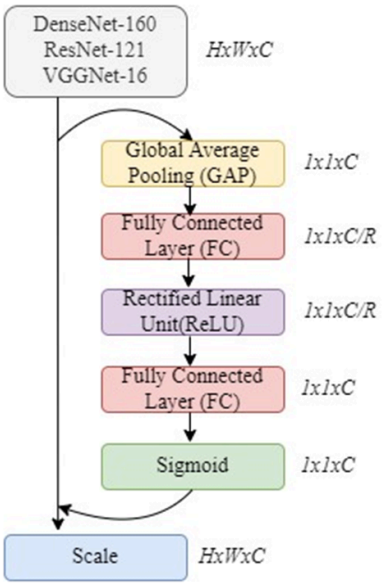

Figure 9. Structure of squeeze-excitation block applied in B2-Net to the sleeted three models (DenseNet-160, ResNet-121, and VGG-16).

Algorithm 1. B2-Net.

Input: IN
Output: OUT
Initialisation:

1. Load the nine ImageNet Model (M_1, M_2….M_9) and set the classification error (C_{err}).
2. Set the model evaluation parameters (P_i) AUROC, accuracy, F1 core, precision, recall, and specificity.
3. Set the hyperparamters (H_1, H_2. H_n) such as epoch size (E_s) dropout rate (D_r) learning rate (L_r).
4. Read the input image (I_1, I_2.I_n) of size 224 × 224.
5. **For** *i* in 0–9: **do**
   Load the *i*th model M_i from model set M.
   Fine tune the hyperparameters H_i
   Train, validate and test the model M_i with H_i.
   Calculate the validation error V_{err}.
   If ($C_{\text{err}} \leqslant V_{\text{err}}$) **then**
   save the new model NM_j
   else:
   go to step 2.
   end if
   end for
6. **For** *i* in 0 to *j*: **do**
   Load the new model NM_j and assign a weight MW_i by using (P_1, P_2.P_6)
   Save the weight W_i of the new model MW_i
   end for
7. Calculate the weighted average of all models W_{avg} using $W_{\text{avg}} = MWi \Big/ \sum_{i=1}^{n} MWi$.
8. **For** *i* in 0–9: **do**
   For *j* in 0–3: **do**
   Load the *i*th model and compare the weighted average W_{avg}.
   If ($W_{\text{avg}}(i) > W_{\text{avg}}(i+1)$) **then**
   save the $i+1$th model
   end for *j*
   Save the *j*th model as the best model.
   end for *i*
9. **For** *j* in 0–3: **do**
   Load the best selected *i*th model.

(Continued.)

Algorithm 1. (Continued.)

Compute the weight W_j of the ith model.
Calculate the ensemble probability using equation (10).
10. **For** i in 0–3: **do**
if ($I_i = +$ve) **then**
$C1 = C1 + W_j$;
else:
$C0 = C0 + W_j$;
end if
end for
11. **If** ($C0 < C1$ && $C0 < C2$) **then**
OUT = Normal Chest x-ray;
else If ($C0 < C1$ && $C1 < C2$) **then**
OUT = Bacterial Pneumonia;
else If ($C0 < C1$ && $C0 > C2$) **then**
OUT = Viral Pneumonia;
Return **OUT**
End

Algorithm 1 explains the ensemble technique applied to the large augmented dataset to classify the chest x-ray as normal, bacterial, or viral pneumonia in the B2-Net network. The nine image classification models $M_1 \ldots M_9$ are initially selected and assigned the classification error. The hyperparameters like epoch size (E_s) and dropout rate (D_r) learning rate (L_r) are used to validate the trained models and select the three best models for fine-tuning and retraining the B2-Net model. The average weight of the models is found using $W_{\text{avg}} = MWi \Big/ \sum_{i=1}^{n} MWi$. In the proposed B2-Net framework, steps 8 and 9 can be replaced by equation (11). An ensemble is formed and trained on the test set utilizing these scores received during the training phase by the nine base models. This method guarantees that the test set is kept separate from the prediction set.

5. Experimental results and discussion

5.1. B2-Net performance evaluation

This research conducted a quantitative and qualitative analysis of the models provided, focusing on their absolute performance. True positive (T_p), true negative (T_n), false positive (F_p), and false negative (F_n) are the four criteria used in qualitative analysis. The most detailed statistic of allaying a model is accuracy, calculated as the percentage of test instances labeled correctly out of the total number of test cases. It can be used for various situations; however, it performs poorly on imbalanced data. Additionally, this work reduces the data imbalance by TL and ensemble. Thus model accuracy is also considered one of the metrics to compare the performance of the B2-Net model. The accuracy of a model for a given dataset is given in equation (16)

$$\text{Accuracy} = \left(T_p + T_n\right) / (T_n + T_p + F_n + F_p). \tag{16}$$

A measure of the correctness of classification is its precision, which is the proportion of true positive classifications to all true positive predictions. Precision in positive class classification indicates a more significant proportion. The model's ability to identify positive class instances is enhanced if the proportion is larger

$$\text{Precision} = ((T_p/(T_p + F_p)). \tag{17}$$

The third criteria for evaluation are recall, often called sensitivity; it is the proportion of true positives among the total number of positives

$$\text{Recall} = ((T_p/(T_p + F_n)). \tag{18}$$

The degree of specificity is defined as the proportion of false-negative results that come from true negative results. It is like recall, but with more emphasis on the negative instances

$$\text{Specificity} = ((T_n/(T_n + F_p)). \tag{19}$$

When both recall and accuracy are equally significant, the F1 score is the best statistic to employ. For a model's F1 score to increase, it has to perform better in terms of accuracy and recall

$$\text{F1 - Score} = 2 \times [(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})]. \tag{20}$$

Table 4. Performance comparison of nine models in normal vs viral pneumonia.

SI. no.	Model	Accuracy	Precision	Recall	Specificity	F1-score	AUROC
1	AlexNet	89.03	81.82	89.06	89.02	85.29	0.9288
2	DenseNet-121	90.36	83.18	91.28	89.86	87.04	0.9501
3	DenseNet-160	97.41	95.92	96.91	97.69	96.41	0.9791
4	MobileNet-V2	90.89	83.81	92.15	90.20	87.78	0.9551
5	MobileNet-V3	91.85	85.31	93.26	91.07	89.11	0.9594
6	ResNet-121	96.11	93.40	95.83	96.26	94.60	0.9896
7	ResNet-152	95.01	91.96	94.33	95.39	93.13	0.9732
8	VGGNet-16	97.04	94.90	96.88	97.13	95.88	0.9916
9	VGGNet-19	94.84	92.00	93.88	95.39	92.93	0.9666

The AU-ROC curve is another important measure for assessing models; it is a plot of the TPR and the FPR for different probability estimates for predicting class labels

$$\text{True Positive Rate (TPR)} = (T_p/T_p + F_p) \tag{21}$$

$$\text{False Positive Rate (FPR)} = (F_p/F_p + T_n). \tag{22}$$

The nine different classifiers' three class (bacteria, viral, normal) classification augmented transfer learning (ATL) strategies are shown in tables 3 and 4, respectively. When performing multi-class classification, the classifier's performance is evaluated based on all three classes' total accuracy and AUROC. The multi-class classification method has an overall AUC of 0.9867 and an accuracy of 95.71% when categorizing viral and bacterial forms of pneumonia. Accuracy, precision, recall, specificity, F1-score, and AUROC curves are used to assess the classification effectiveness of models. In tables 3 and 4, the highlighted models DenseNet-160, ResNet-121, and VGGNet-16 have achieved better performance in ATL approach. Some other models like MobileNet-V2, DenseNet-121, and VGG-19 have performed better in a few parameters, and it has reflected as one of the best models in the ATL experiment.

The AUROC curve is a summary metric describing a model's performance over several thresholds. Noteworthy is the fact that an AUROC score of one represents a perfect score, while an AUROC value of 0.5 represents guessing at random. The AUROC curve of the nine models used in this work is plotted in figures 10 and 11.

In figures 10 and 11, nine traditional ML models are compared. ResNet-160, VGGNet-16, and VGGNet-19 have done exceptionally well in bacterial pneumonia identification, and for the case of viral pneumonia detection, ResNet-160, DenseNet-152 a, VGGNet-16 have led with maximum accuracy and AUROC curves. From the study, as mentioned earlier, developers decided and selected the top three models for future exploration. Since it is a supervised learning method, the confusion matrix is one of the most accurate metrics to evaluate its effectiveness. A confusion matrix is a method for neatly mapping predictions back to the initial classes from which the data came. The confusion matrix obtained from the two-class classification model is shown in figures 12(a)–(e).

Table 5 summarizes the comparison of B2-Net models in six selection parameters. The primary purpose of this study is to create a diagnostic tool powered by GPUs that can recognize and categorize pneumonia in chest x-rays. Testing TL ensemble techniques on a balanced dataset revealed that DenseNet-160, ResNet-121, and VGGNet-16 all obtained AUCs of 0.9801, 0.9822, 0.9955, respectively and the proposed B2-Net approach outperforms all other three by 0.9977 AUROC.

Figures 13(a) and (b) show the classification efficiency of the top three and B2-Net models, respectively. The above models are employed in a three-class classification mode to categorize normal, viral, and bacterial pneumonia from chest x-ray images. An actual label and a forecasted label are shown in the confusion matrix. Prediction accuracy for the VP, BP, and normal chest x-ray datasets are shown in the diagonal columns. Compared to the data in table 5, the B2-Net strategy outperforms all three models used throughout the research duration in AUROC and the other five evaluation matrices. The AUROC for the proposed B2-Net ensemble methodology for pneumonia classification using DL methods is shown in figure 14.

The augmented x-ray images are trained and tested on the high-performance computing platform to reduce development time and cost. The suggested B2-Net strategy surpasses all the models in the AUROC curve, precision, recall, and F1 score, as can be shown by comparing the AUCROC curve for the six contemporary models utilized in this ATL methodology shown in figures 10 and 11. Figure 11 shows an increase in performance of up to 0.1% and 0.3% for DenseNet-160 and VGGNet-16, respectively; the

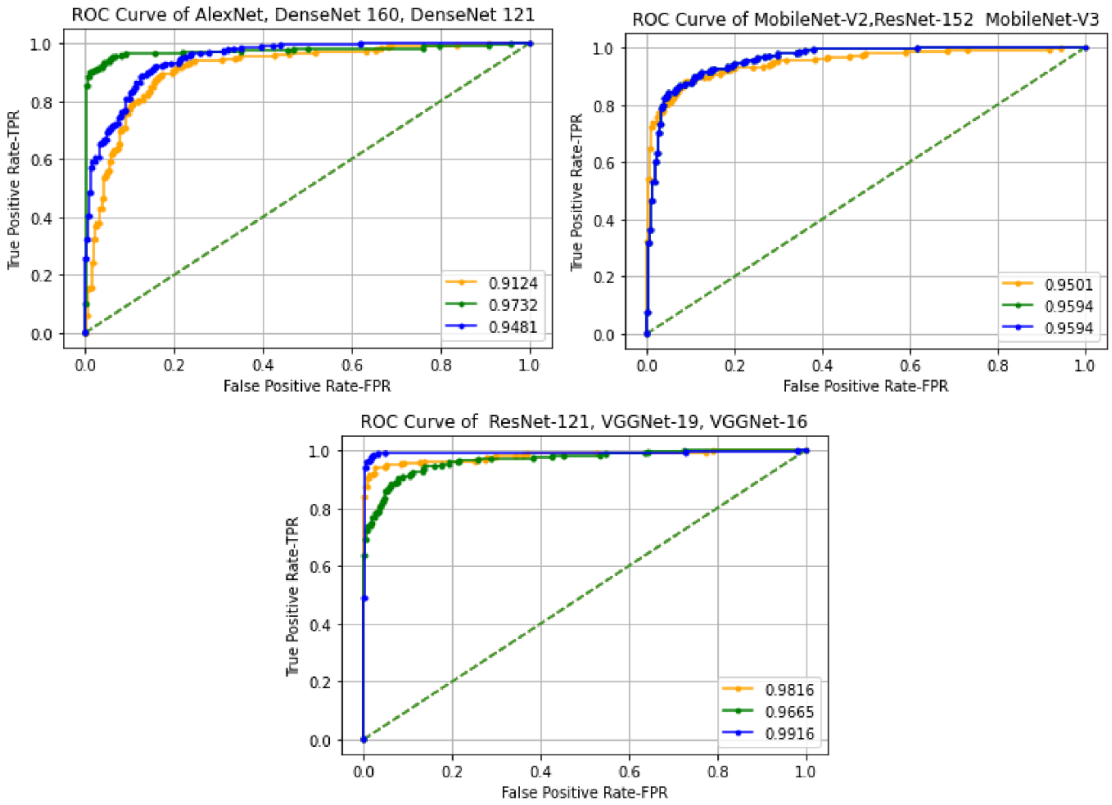

Figure 10. AUROC metric of bacterial pneumonia framework computed for (a) AlexNet, DenseNet 121, DenseNet 160, (b) MobileNet-V2, MobileNet-V3, ResNet-121, (c) ResNet-152 VGGNet-16, VGGNet-19.

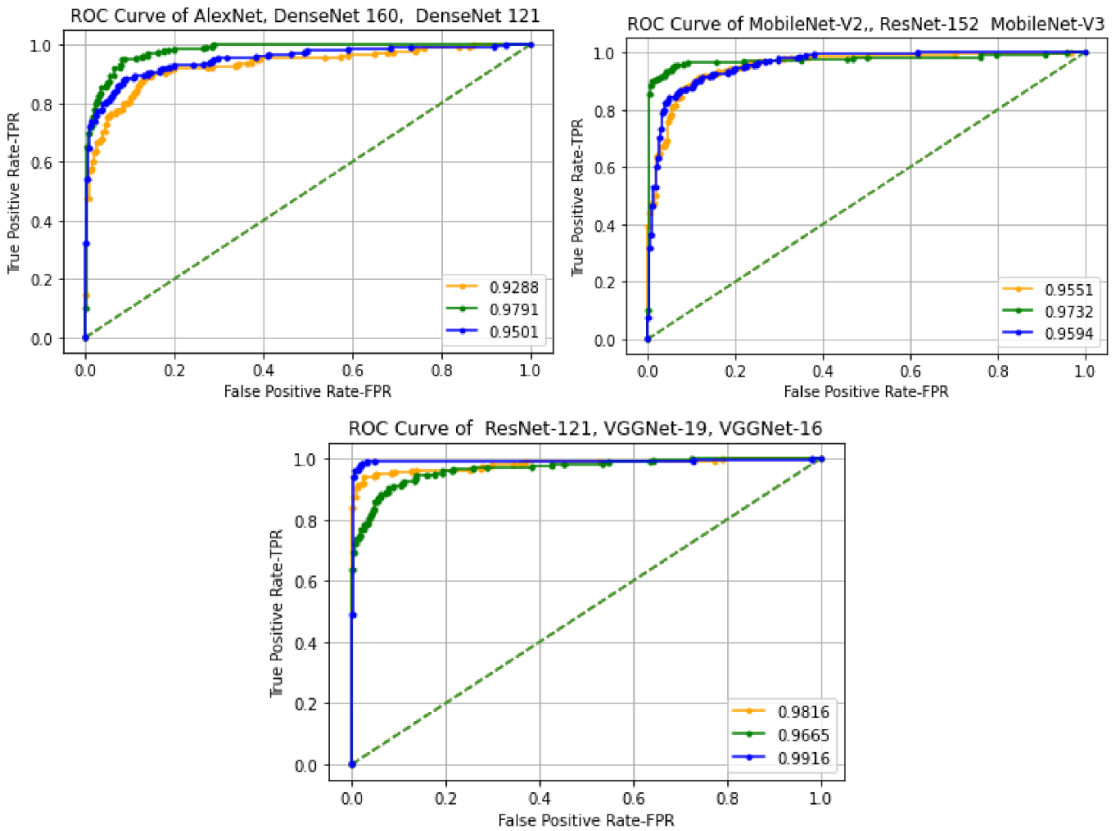

Figure 11. AUROC metric of viral pneumonia framework computed for (a) AlexNet, DenseNet 121, DenseNet 160, (b) MobileNet-V2, MobileNet-V3, ResNet-121, (c) ResNet-152 VGGNet-16, VGGNet-19.

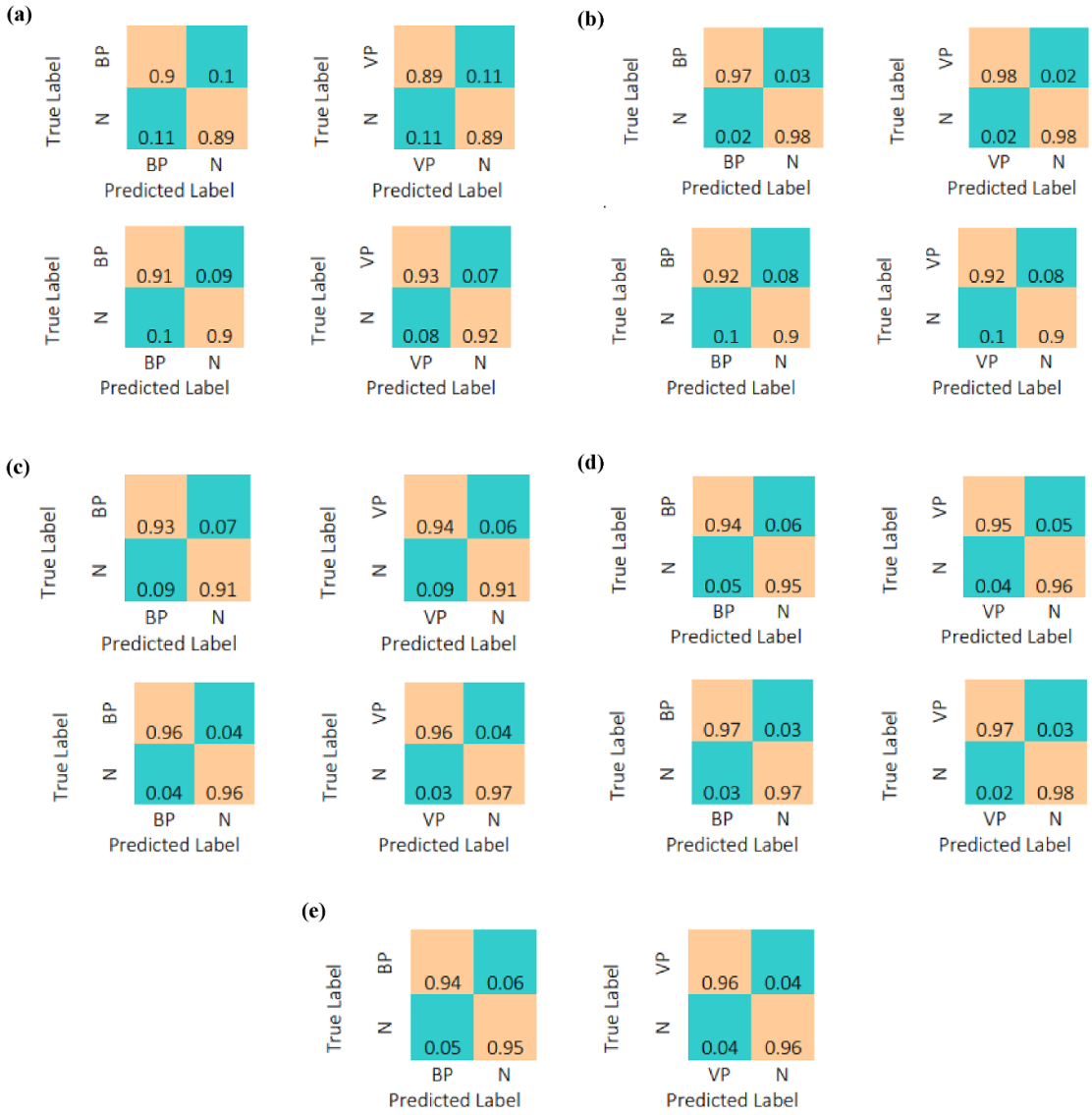

Figure 12. (a) Confusion matrix of AlexNet and DenseNet-121 model for bacterial and viral pneumonia detection from chest x-ray. 'N' represents normal images, 'BP' and 'VP' represents bacterial and viral pneumonia respectively. (b) Confusion matrix of DenseNet 160 and MobileNet-V2 model for bacterial and viral pneumonia detection from chest x-ray. (c) Confusion matrix of MobileNet-V3 and ResNet-121 model for bacterial and viral pneumonia detection from chest x-ray. (d) Confusion matrix of ResNet-152 and VGGNet-16 model for bacterial and viral pneumonia detection from chest x-ray. (e) Confusion matrix of VGGNet-19 model for bacterial and viral pneumonia detection from chest x-ray.

Table 5. Performance comparison of B2-Net models in accuracy, sensitivity, specificity, and F1-score metrics.

SI. no.	Model	Accuracy	Precision	Recall	Specificity	F1-score	AUROC
1	DenseNet-160	97.21	95.48	94.87	98.17	95.18	0.9801
2	ResNet-121	96.10	93.55	92.95	97.38	93.25	0.9822
3	VGGNet-16	98.33	98.04	96.15	99.21	97.09	0.9955
4	B2-Net	98.88	98.08	98.08	99.21	98.08	0.9977

performance of ResNet-121 is degraded up to 0.0745%. The validation and training accuracies of the best three models using the B2-Net ensemble approach are displayed in figures 15–17.

The accuracy of the DenseNet-160 model during training and validation utilizing the TL-enabled ensemble architecture is shown in figure 15. The model is 85.54% accurate during training and 84.70% during validation. This model's optimum degree of accuracy is achieved by deploying the GPU system throughout the training and validation procedures; however, the total number of epochs is limited to 100.

Figure 16 demonstrates that the ResNet-121 model is more accurate during training and validation than the DenseNet-160 model. B2-Net-enabled ResNet-121 achieves 85.81% training accuracy and 84.6%

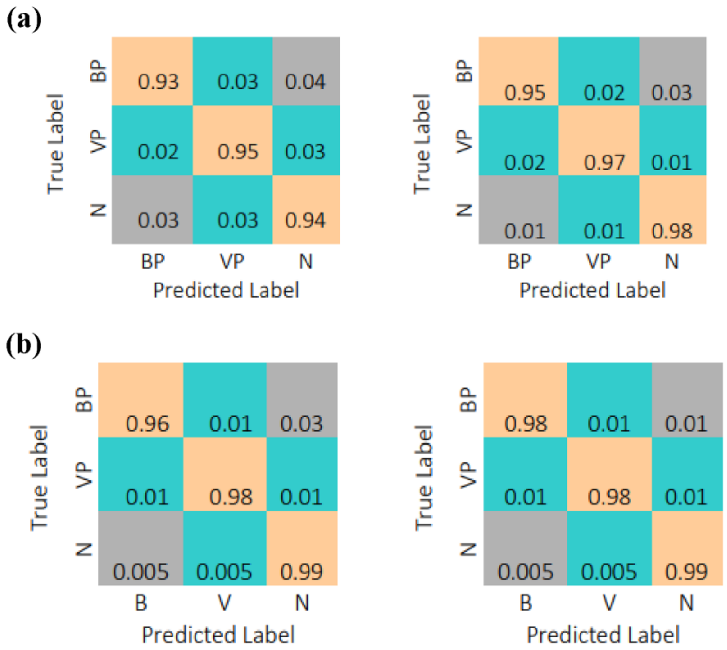

Figure 13. (a) Confusion matrix of DenseNet-160 and ResNet-121 model for normal, bacterial and viral pneumonia detection from chest x-ray. (b) Confusion matrix of VGGNet-16 models and proposed B2-Net ensemble technique; 'N' represents normal images, 'BP' and 'VP' represents bacterial and viral pneumonia respectively.

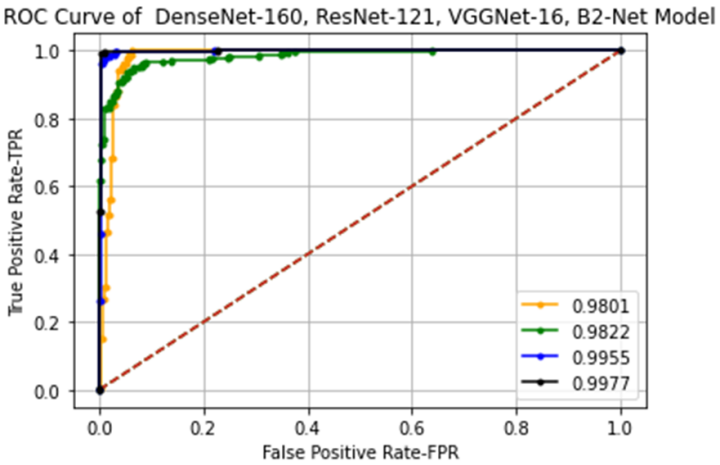

Figure 14. AUROC metric computed for the ensemble technique on DenseNet-160, ResNet-121, VGGNet-16 models and proposed B2-Net ensemble technique.

validation accuracy. The model accuracy graph also depicts the evolution of the training and validation accuracies over time, demonstrating that a better learning rate is attained after the 30th epoch.

This research employs a uniform number of epochs across all models to standardize the model selection procedure and decrease variability. With a training accuracy of 89.35% and a validation accuracy of 88.51, the VGGNet-16 model performs better than the abovementioned models. Tables 4 and 5 show that VGGNet-16 is the clear front-runner of this investigation.

Figure 18 highlights that when the three best models are blended, the resulting model performs very well in training and validation. The proposed B2-Net method allows researchers to achieve 90.06% accuracy during training and 88.87% during validation. This model's learning rate is a crucial hyper-parameter; a lower learning rate is selected, requiring longer training and validation epochs. The Adam optimizer [27] enhanced accuracy at the ensemble dataset with a learning rate of 0.0001. The learning rate in the second and third ensemble models is optimal after the 40th epoch. The proposed B2-Net method surpasses all three models in terms of accuracy and learning epochs since it has a constant rate from the 30th epoch. Compared to the previous three models, the model performance is much improved by using TL and ensemble

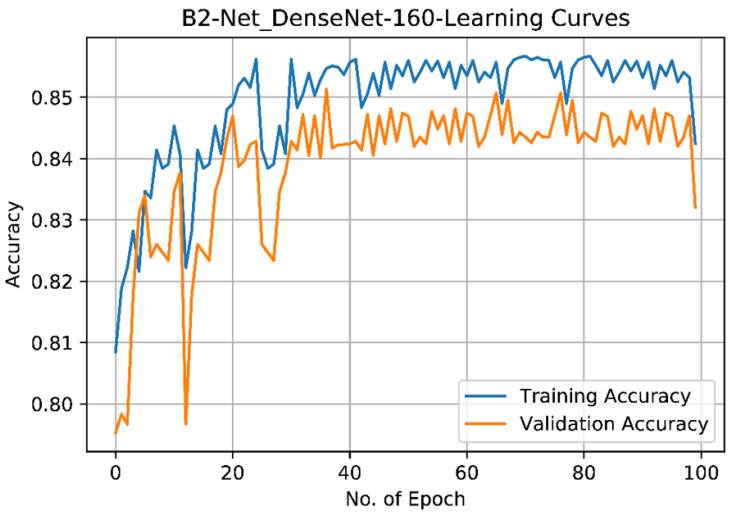

Figure 15. Training and validation accuracy of proposed B2-Net approach on DenseNet-160 model.

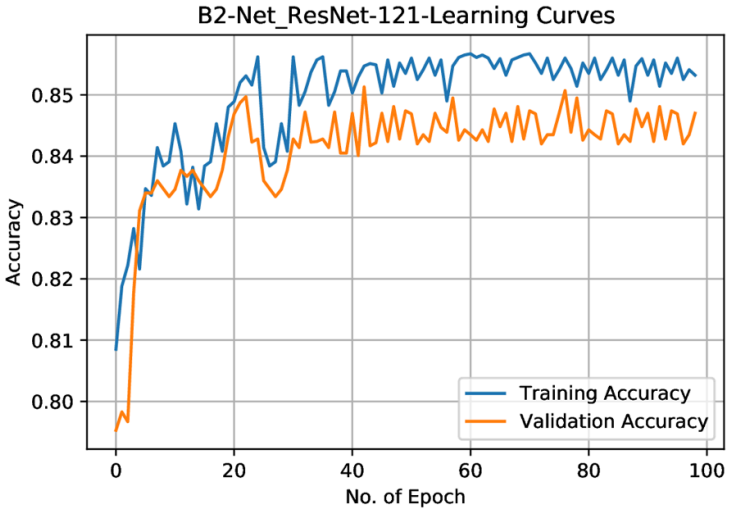

Figure 16. Training and validation accuracy of proposed B2-Net approach on ResNet-121.

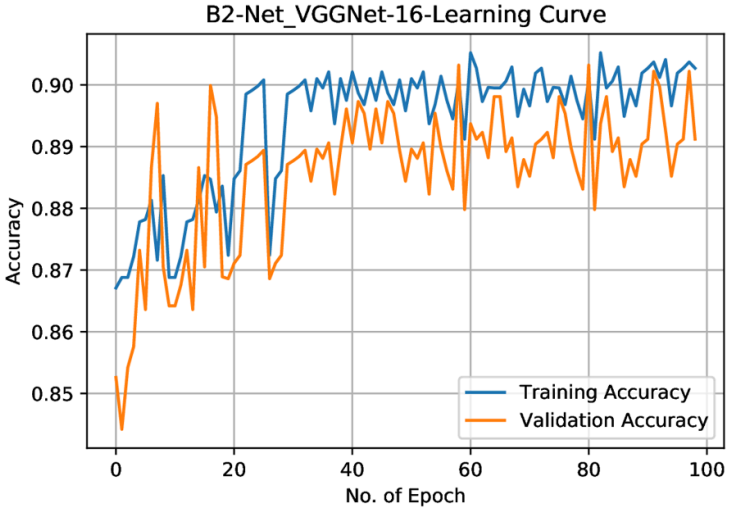

Figure 17. Training and validation accuracy of proposed B2-Net approach on VGGNet-16.

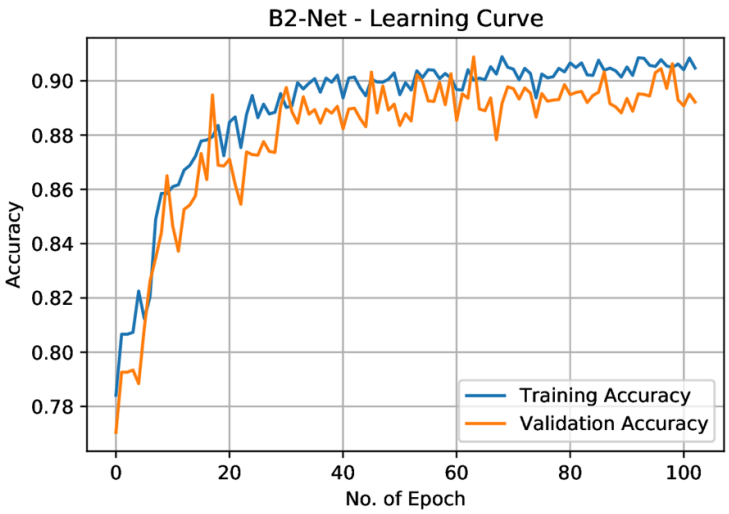

Figure 18. Training and validation accuracy of proposed B2-Net ensemble learning method.

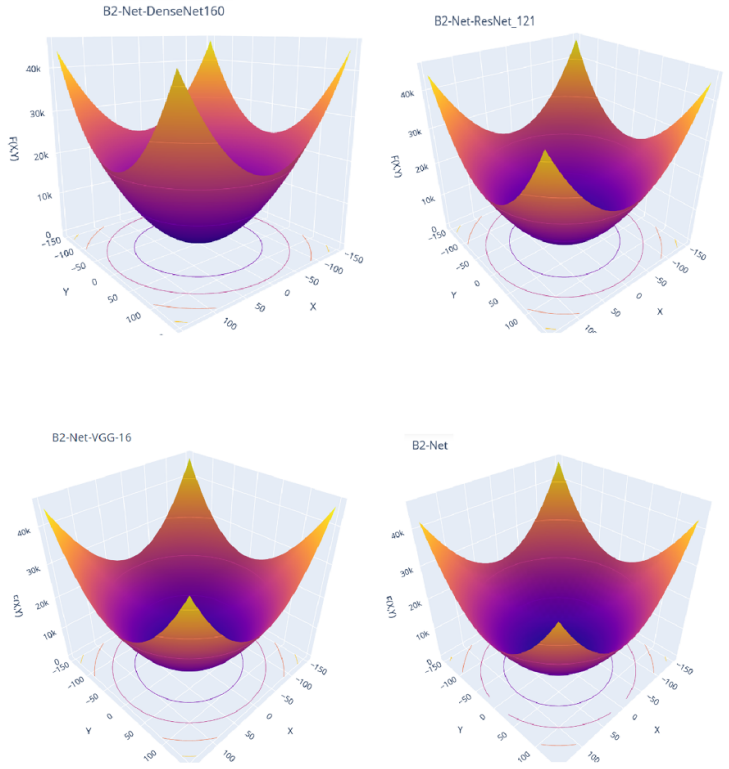

Figure 19. Convex loss landscape of DenseNet-160, ResNet-121, VGG-16 and B2-Net models.

techniques. After 40 epochs of retraining, the accuracy of the three models remains almost constant, and the model overfitting issue is decreasing owing to the integration of the ensemble technique.

The loss function in a neural network depends on several factors, including the model's design, optimization strategy, initialization, etc. For each configuration, the resultant loss function is represented with the Y axis displaying the loss function and the X axis displaying the number of epochs. In this part, some visual representations are used to investigate the structure of neural loss functions and the impact of loss landscapes on the B2-Net model. The convex loss landscape of the trained B2-Net models is shown in figure 19.

The minimal optimal solution is highly generalizable, and landscape visualization provides deeper insights into how and why neural networks can optimize even the most complicated non-convex functions. It is well-known that network architectural designs that use skip connections, like ResNet, result in loss

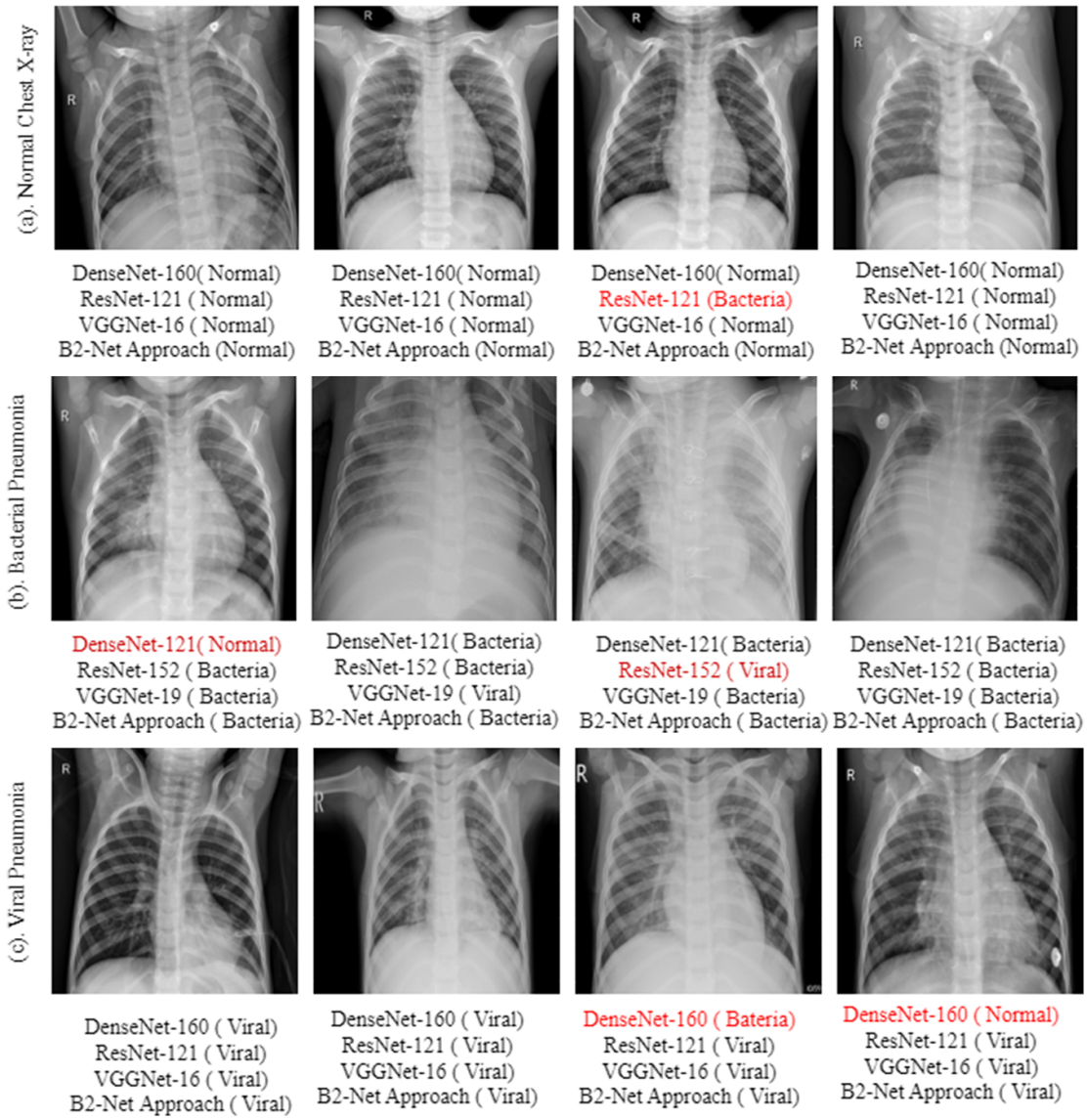

Figure 20. Chest-x-ray samples of (a) healthy patient (b) bacterial pneumonia and (c) viral pneumonia diagnosed and misdiagnosed by the proposed B2-Net ensemble techniques.

functions that train more smoothly. As seen in the figure, the B2-Net model, the B-Net rennet 121, and VGG16 frameworks, in particular, benefit from minimizers trained with well-selected parameters such as batch size, learning rate, and optimizer.

5.2. Implementation details

NVIDIA Jetson Nano GPU hardware is used to train, test and validate the six different CNN models that had previously been pre-trained. The Jetson Nano is a compact, high-performance computing module used in embedded systems, AI, computer vision systems, and internet of things environments. This GPU module is driven by a Quad-core ARM Cortex-A57 MPCore CPU and 4 GB of 64 bit LPDDR4 memory. Experiments are run on the top three models, with the results being analyzed in a TensorFlow framework. TensorFlow, developed by Google, is a competitive framework that can train and execute ML, DL, DNNs, and NLP.

Figure 20 shows three groups of x-ray images with the matching test results for the three TL ensemble models and the proposed B2-Net model employed in this research. DenseNet-160 misidentified three test images as bacterial and normal chest x-rays. VGGNet-16 and ResNet-121 models misdiagnosed one x-ray image each. The B2-Net and the other three models are evaluated and validated using 657 and 538 chest x-ray images from the pooled dataset, as shown in table 2. Although the false negative rate is the most critical criterion for a medical image classifier, rather than accuracy and other factors, the proposed B2-Net model is the best alternative for real-time clinical applications. This approach, particularly during the present worldwide Covid'19 outbreaks, could assist radiologists in swiftly and reliably diagnosing pneumonia,

allowing for early treatment and prevention of patients. However, the proposed B2-Net approach outperforms all other models and scores a 100% recall rate. Finally, throughout the validation phase, all of the B2-Net models performed well in terms of diagnostic accuracy.

6. Conclusion

Medical imaging is one of the fastest and most complicated forms of patient data. To get previously unavailable insights into clinical decision-making and to derive meaning from unstructured data assets, AI, and ML have captured the attention of the healthcare sector. Creating efficient models to analyze large numbers of medical images without false negatives to diagnose illness is the most need of the current healthcare industry. The developed B2-Net framework employs a DWC, squeeze-excitation, TL, DA, and weighted majority vote ensemble classification strategy to train the nine top ImageNet-trained models (AlexNet, DenseNet-121, DenseNet-160, MobileNet-V2, MobileNet-V3, ResNet-121, ResNet-152, VGGNet-16, and VGGNet-19). Three of the nine best models (DenseNet-160, ResNet-121, and VGGNet-16) are chosen based on the research performance assessment criteria and retrained for the new model. An ensemble approach and DWCs are then used for the retrained and best-selected models, which outperform all other state-of-the-art models employed in this research. The suggested B2-Net framework distinguishes between normal, bacterial, and viral pneumonia in chest x-ray images with an astounding 97.69% accuracy, 100% recall, and 0.9977 AUROC values. The research also discovered that performance could be enhanced further by expanding the size of the dataset using a DA and TL methodology and by improving classification using an ensemble method. Thorough investigations will expand to classify more lung disorders, including TB and Covid'19 variants.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

K M Abubeker https://orcid.org/0000-0001-7646-0781
S Baskar https://orcid.org/0000-0003-3570-3059

References

[1] Chowdhury M E H *et al* 2020 Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* **8** 132665–76
[2] Chouhan V, Singh S K, Khamparia A, Gupta D, Tiwari P, Moreira C, Damaševičius R and de Albuquerque V H C 2020 A novel transfer learning based approach for pneumonia detection in chest x-ray images *Appl. Sci.* **10** 559
[3] Bushara A R and Vinod Kumar R S 2022 Deep learning-based lung cancer classification of CT images using augmented convolutional neural networks *Electron. Lett. Comput. Vis. Image Anal.* **21** 130–42
[4] Halder A and Datta B 2021 *Mach. Learn.: Sci. Technol.* **2** 045013
[5] Yao S, Chen Y, Tian X and Jiang R 2021 Pneumonia detection using an improved algorithm based on faster R-CNN *Comput. Math. Methods Med.* **2021** 8854892
[6] Rahman T, Chowdhury M E H, Khandakar A, Islam K R, Islam K F, Mahbub Z B, Kadir M A and Kashem S 2020 Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest x-ray *Appl. Sci.* **10** 3233
[7] Alqudah A and Alqudah A M 2021 Sliding window based deep ensemble system for breast cancer classification *J. Med. Eng. Technol.* **45** 313–23
[8] Oyelade O N, Ezugwu A E and Chiroma H 2021 CovFrameNet: an enhanced deep learning framework for pneumonia detection *IEEE Access* **9** 77905–19
[9] Lacruz F and Vidarte R 2020 Analysis of deep learning models for pneumonia diagnosis from x-ray chest images *Researchgate* **2** 127–35
[10] Mehmood S, Ghazal T M, Khan M A, Zubair M, Naseem M T, Faiz T and Ahmad M 2022 Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing *IEEE Access* **10** 25657–68
[11] Kotei E and Thirunavukarasu R 2022 Ensemble technique coupled with deep transfer learning framework for automatic detection of tuberculosis from chest x-ray radiographs *Healthcare* **10** 2335
[12] Kundu R, Das R, Geem Z W, Han G-T and Sarkar R 2021 Pneumonia detection in chest X-ray images using an ensemble of deep learning models *PLoS One* **16** e0256630
[13] Liz H, Sánchez-Montañés M, Tagarro A, Domínguez-Rodríguez S, Dagan R and Camacho D 2021 Ensembles of convolutional neural network models for pediatric pneumonia diagnosis *Future Gener. Comput. Syst.* **122** 220–33
[14] Fraiwan L, Hassanin O, Fraiwan M, Khassawneh B, Ibnian A M and Alkhodari M 2021 Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers *Biocybern. Biomed. Eng.* **41** 1–14
[15] Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning *J. Big Data* **6** 1–48

[16] Liu D, Liu J, Yuan P and Feng Y 2022 A data augmentation method for prohibited item x-ray pseudocolor images in x-ray security inspection based on Wasserstein generative adversarial network and spatial-and-channel attention block *Comput. Intell. Neurosci.* **2022** 8172466
[17] Sharma H, Jain J, Bansal P and Gupta S Feature extraction and classification of chest x-ray images using CNN to detect pneumonia *2020 10th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)* pp 227–31
[18] Stephen O, Sain M, Maduh U and Jeong D 2019 An efficient deep learning approach to pneumonia classification in healthcare *J. Healthc. Eng.* **2019** 1–7
[19] Motamed S and Rogalla P 2021 Data augmentation using generative adversarial networks (GANs) for GAN-based detection of pneumonia and COVID-19 in chest x-ray images *Inform. Med. Unlocked* **27** 1–7
[20] Zou L 2023 Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections *IEEE Trans. Artif. Intell.* **4** 242–54
[21] Ayan E, Karabulut B and Ünver H M 2022 Diagnosis of pediatric pneumonia with ensemble of deep convolutional neural networks in chest x-ray images *Arab. J. Sci. Eng.* **47** 2123–39
[22] Li K, Zheng F, Wu P, Wang Q, Liang G and Jiang L 2022 Improving pneumonia classification and lesion detection using spatial attention superposition and multilayer feature fusion *Electronics* **11** 3102
[23] Hussain L, Nguyen T, Li H, Abbasi A A, Lone K J, Zhao Z, Zaib M, Chen A and Duong T Q 2020 Machine-learning classification of texture features of portable chest x-ray accurately classifies COVID-19 lung infection *Biomed. Eng. Online* **19** 88
[24] Theodoridis T, Loumponias K, Vretos N and Daras P 2021 Zernike pooling: generalizing average pooling using Zernike moments *IEEE Access* **9** 121128–36
[25] Deng J, Dong W, Socher R, Li L-J, Li K and Fei-Fei L 2009 ImageNet: a large-scale hierarchical image database *2009 IEEE Conf. on Computer Vision and Pattern Recognition (Miami, FL, USA)* pp 248–55
[26] Kaggle Chest x-ray images (pneumonia) (available at: www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia) (Accessed 9 November 2022)
[27] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)