Kusumadhara S, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 5, (Series-I) May 2021, pp. 61-73

RESEARCH ARTICLE OPEN ACCESS

A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of MRMR Feature Selection and Machine Learning Technique

[1]Kusumadhara S, [2]Nishmitha K, [3]Pavithra S, [4]Pooja H J, [5]Pratheeka A P
[1]Associate Professor, Department of E&C, KVGCE, Sullia, [2345]Final Year B.E, KVGCE, Sullia

ABSTRACT – Pneumonia causes the death of around 700,000 children every year and affects 7% of the global population. Chest X-rays are primarily used for the diagnosis of this disease. However, even for a trained radiologist, it is a challenging task to examine chest X-rays. There is a need to improve the diagnosis accuracy. In this work, an efficient model for the detection of pneumonia trained on digital chest X-ray images is proposed, which could aid the radiologists in their decision making process. A novel approach based on a weighted classifier is introduced, which combines the weighted predictions from the state-of-the-art deep learning models such as Convolution Neural Network, ResNet18, Xception, InceptionV3, DenseNet121, and MobileNetV3 in an optimal way. This approach is a supervised learning approach in which the network predicts the result based on the quality of the dataset used. Transfer learning is used to fine-tune the deep learning models to obtain higher training and validation accuracy. Partial data augmentation techniques are employed to increase the training dataset in a balanced way. The proposed weighted classifier is able to outperform all the individual models. Finally, the model is evaluated, not only in terms of test accuracy, but also in the AUC score.
The final proposed weighted classifier model is able to achieve a test accuracy of 98.43% and an AUC score of 99.76 on the unseen data from the Guangzhou Women and Children's Medical Centre pneumonia dataset. Hence, the proposed model can be used for a quick diagnosis of pneumonia and can aid the radiologists in the diagnosis process.

KEYWORDS: Pneumonia; Chest X-ray Images; Convolution Neural Network (CNN); Deep Learning; Transfer Learning; Computer-aided Diagnostics.

I. INTRODUCTION

Pneumonia is an acute respiratory infection that affects the lungs. It is a fatal illness in which the air sacs get filled with pus and other liquid [1]. There are mainly two types of pneumonia: bacterial and viral. Generally, it is observed that bacterial pneumonia causes more acute symptoms. The most significant difference between bacterial and viral pneumonia is the treatment. Treatment of bacterial pneumonia is done using antibiotic therapy, while viral pneumonia will usually get better on its own [2]. It is a prevalent disease all across the globe. Its principal cause includes a high level of pollution. Pneumonia is ranked eighth in the list of the top 10 causes of death in the United States [3]. Due to pneumonia, every year, 3.7 lakh children die in India, which constitutes a total of fifty percent of the pneumonia deaths that occur in India [4]. The disease frequently goes overlooked and untreated until it has reached a fatal point, especially in the case of old patients. It is the single largest cause of death in children (especially under the age of five) worldwide [5]. According to the WHO, "Every year, it kills an estimated 1.4 million children under the age of five years, accounting for 18% of all deaths of children under five years old worldwide. Pneumonia affects children and families everywhere but is most prevalent in South Asia and sub-Saharan Africa. Children can be protected from pneumonia. It can be prevented with simple interventions and treated with low-cost, low-tech medication and care" [2]. Therefore, there is an urgent need to do research and development on computer-aided diagnosis so that the pneumonia-related mortality, especially in children, can be reduced.

One of the following tests can be done for pneumonia diagnosis: chest X-rays, CT of the lungs, ultrasound of the chest, needle biopsy of the lung, and MRI of the chest [6]. Currently, chest X rays are one of the best methods for the detection of pneumonia [7]. X-ray imaging is preferred over CT imaging because CT imaging typically takes considerably more time than X-ray imaging, and sufficient high-quality CT scanners may not be available in many underdeveloped regions. In contrast, X-rays are the most common and widely available diagnostic imaging technique, playing a crucial role in clinical care and epidemiological studies [8, 9].

There are several regions across the globe where there is a scarce availability of practiced healthcare workers and radiologists whose prediction on such diseases matter greatly [10–12]. Computer aided diagnosis using artificial intelligence based solutions is becoming increasingly popular these days [13, 14].

This facility can be made available to a large population at a minimal cost. Another issue with this disease is that sometimes, the features that describe the very existence of the disease often get mixed with other diseases, and hence, radiologists find it challenging to diagnose this disease. Deep learning techniques solve all these problems, and their accuracy in the prediction of the disease is the same and sometimes even greater than an average radiologist [15]. Among the deep learning techniques, convolutional neural networks (CNNs) have shown great promise in image classification and segmentation and therefore are widely adopted by the research community. Biomedical image diagnosis that uses the techniques of deep learning and computer vision has proven to be very helpful to provide a quick and accurate diagnosis of the disease that matches the accuracy of a reliable radiologist [16]. Currently, deep learning based methods cannot replace trained clinicians in medical diagnosis, and they aim to supplement clinical decision making. In this project, a model is presented based on the applications of deep learning and convolutional neural networks that are capable of classifying automatically that the patient has pneumonia or not. The proposed methodology uses a deep transfer learning algorithm that extracts the features from the X-ray image that describes the presence of disease automatically and reports whether it is a case of pneumonia.

II. LITERATURE SURVEY

Deep learning based methods are already being used in various fields [17–21]. Different authors have already proposed several biomedical image detection techniques. M.I.Razaak [22] discussed the challenges and the future of medical image processing. Much work has already been done for the detection of numerous diseases by using deep learning based techniques, as stated by Dinggang Shen [23].

Andre [24] presented a deep learning model for dermatologist-level classification of skin cancer, and F. Milletari [25] also proposed a methodology for the depiction of prostrate in MRI volumes using CNN. Grewal [26] used the technique of deep learning for brain hemorrhage detection in CT scans, and Varun [27] proposed a method for detecting diabetic retinopathy in retinal fundus photographs. Y. Bar [28] also discussed chest pathology detection by the techniques based on deep learning. Methods regarding the examination of the detection of disease by chest X-ray have also been worked on earlier by performing various examination techniques [29–31]. The chest X-ray images are passed through the evaluation process of scan line optimization such that it eliminates all the other body parts to avoid any error in diagnosis. The algorithm was described by S. Hermann [32]. Nasrullah et al. [33] used two deep three-dimensional (3D) customized mixed link network (CMixNet) architectures for lung nodule detection and classification.

Yao L et al. [34] combined DenseNet and long-short term memory networks (LSTM) to exploit the dependencies between abnormalities. Several authors also have worked on pneumonia classification. Khatri et al. [35] proposed the use of EMD (earth mover's distance) to identify infected pneumonia lungs from normal non-infected lungs. Rahib et al. [36] and Okeke et al. [37] used a CNN model for pneumonia classification. Some researchers have shown assuring results such as Cohen et al. [38] and Rajaraman et al. [39]. Rajaraman et al. [39] tried to explain the performance of customized CNNs to detect pneumonia and further differentiate between bacterial and viral types in pediatric CXRs

Sirazitdinov et al. [40] used a region based convolutional neural network for segmenting the pulmonary images along with image augmentation for pneumonia identification. Lakhani and Sundaram [41] used the AlexNet and GoogLeNet neural networks with data augmentation and without any pre-training to obtain an area under the curve (AUC) of 0.94–0.95. Rajpurkar et al. [42] used CheXNeXt, a very deep CNN with 121 layers, to detect 14 different pathologies, including pneumonia, in frontal-view chest X-rays. A localization approach based on pre-trained DenseNet-121, along with feature extraction, was used to identify 14 thoracic diseases in [43]. Saraiva et al. [44], Ayan et al. [45], and Rahman et al. [46] used deep learning based methods for pneumonia classification. Xiao et al. [47] proposed a novel multi-scale heterogeneous three dimensional (3D) convolutional neural network (MSH-CNN) based on chest computed tomography (CT) images.

Xu et al. [48] used a hierarchical convolutional neural network (CNN) structure and a novel loss function, sin-loss, for pneumonia detection. Jaiswal et al. [49] used Mask-RCNN, utilizing both global and local features for pulmonary image segmentation, with dropout and L2 regularization, for pneumonia identification. Jung et al. [50] used a 3D deep CNN (3D DCNN), which had shortcut connections. Vikash et al. [51] combined the outputs of different neural networks and reached the

final prediction using majority voting. None of the above-mentioned approaches except that of Vikash et al. [51] tried to combine predictions from different neural networks.

The main contribution is a weighted classifier that integrates five deep learning models. The weights for each model are based on each model's performance on the testing dataset. This paper is structured as follows: Section 3 deals with the methods used. A brief description of the methods used in this paper is given. The experimental dataset is introduced in Section 4. In Section 5, and the proposed methodology is discussed. In Section 6, the results obtained are discussed concerning different parameters. This section is followed by Section 7, containing the discussion, and Section 8, containing the conclusion of the project.

III. PROBLEM DEFINATION

In recent time, exploration of Machine learning (ML) algorithms in detecting thoracic diseases has gained attention in research area of medical image classification. Lakhani and Sundaram (2017) [12] proposed a method of detecting pulmonary tuberculosis following the architecture of two different DCNNs AlexNet and GoogleNet. Lung nodule classification mainly for diagnosing lung cancer proposed by Huang et al. [13] also adopted deep learning techniques. Performance of different variants of Convolutional Neural Networks (CNNs) for abnormality detection in chest X-Rays was proposed by Islam et al. [14] using the publicly available OpenI dataset [15]. For the better exploration of machine learning in chest screening, Wang et al. (2017) [16] released a larger dataset of frontal chest X-Rays.

Recently, Pranav Rajpurkar, Jeremy Irvin, et al. (2017) [17] explored this dataset for detecting pneumonia at a level better than radiologists, they referred their model as ChexNet which uses DenseNet-121 layer architecture for detecting all the 14 diseases from a lot of 112,200 images available in the dataset. After the CheXNet[17] model, Benjamin Antin et al.(2017) [18] worked on the same dataset and proposed a logistic regression model for detecting pneumonia. Pulkit Kumar, Monika Grewal (2017) [19] using the cascading convolutional networks contributed their research for multilabel classification of thoracic diseases. Zhe Li (2018) [20] recently proposed a convolutional network model for disease identification and localization.

IV. PROPOSED SYSTEM

This section deals with the detailed description of the applied methodology. The proposed pneumonia detection system using the 'Densely Connected Convolutional Neural Network' (DenseNet-169) is described in Figure 2. The architecture of the proposed model has been divided into three different stages - the preprocessing stage, the feature extraction stage and the classification stage.

4.1: PRE – PROCESSING STAGE

The primary goal of using Convolutional Neural Network in most of the image classification tasks is to reduce the computational complexity of the model which is likely to increase if the input are images. The original 3-channel images were resized from 1024×1024 into 224×224 pixels to reduce the heavy computation and for faster processing. All of the further techniques has been applied over these downsized images.

4.2: THE FEATURE EXTRACTION STAGE

Although, the features were extracted with different variants of pre-trained CNN models the statistical results obtained proposed DenseNet-169 as the optimal model for the feature extraction stage. Therefore, this stage deals with the description of DenseNet-169 model architecture and its contribution in feature extraction.

The layers between these dense blocks are referred to as transition layers. Each transition layer in the network consists of a batch normalization layer and an 1×1 convolutional layer follow edbya 2×2 average pooling layer that uses a stride of 2. As mentioned above there are 4 dense blocks, each of which contains 2 convolution layers first is of size 1 × 1 followed by 3×3. The size of all the four dense blocks in DenseNet169 architecture pretrained on ImageNet is 6, 12, 32 and 32. Next to this is the final layer that is the classification layer which performs the global average pooling of 7×7 followed by a final fully-connected layer which uses 'softmax' as the activation.

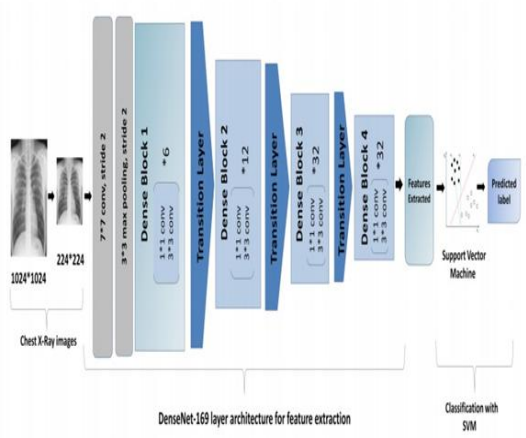

FIG 1. REPRESENT A FLOW DIAGRAM OF OUR METHODOLOGY APPLIED

Architecture of DenseNet 169, Deep Convolutional Networks (DCNNs) have become the most productive frameworks for image recognition because of the presence of peculiar types of the convolutional and pooling layers. But as the network gets deeper the input information or gradient passing through most of the layers gets vanished by the time the last layer of the network is reached. DenseNets overcome this problem of gradient vanishing by connecting all the layers with equal featutre-sizes directly with each other. The chief motive of using DenseNet architecture as a feature extractor is that deeper the network more generic features can be obtained. The pre-trained Densely Connected Convolutional Neural Network of 169 layers (DenseNet-169) has been used for the feature extraction process. This model was proposed by Huang et al. (2016) [9] and the variant used in our study is trained on the large scale publicly available ImageNet dataset. The DenseNet169 architecture comprises of one convolution and pooling layer at the beginning, 3 transition layers, and 4 dense blocks.

After these layers, the final layer i.e the classification layer is present. The first convolutional layer performs 7×7 convolutions with stride 2 followed by a max pooling of 3×3 used with stride 2. Then the network consists of a dense block followed by 3 sets each of which consist a transition layer followed by a dense block. The dense connectivity as proposed by Huang et.al [9] in DenseNets are received by bringing in direct connections from any layer to any other layer in the network. The lst layer in the network receives the feature-maps of all the preceding layers thus ameliorating the flow of gradient throughout the entire network. This requires the concatenation of the feature-maps of the preceding layers which cannot be done unless all the feature-maps are of the same sizes but as the Convolutional Neural Networks primarily intend towards the down sampling of size of feature-maps, the DenseNets architecture is divided into multiple densely connected dense blocks mentioned above.

4.3: EXTRACTION OF FEATURES

The process of feature extraction from the model explained in this section 4.2.1 applies all the layers of the network except the final classification layer. The final feature representation obtained were interpreted as a 50176×1 dimension vector which then supplied as input to different classifiers.

V. SOFTWARE REQUIREMENT SPECIFICATION

A. Software Requirements

- Operating System: Windows 8.1 Platform or Above
- Programming Language: Python 3.6.7
- Framework: Jupiter Notebook
- Cloud Platform: Google Cloud Engine (GCE)

B. Hardware Requirements

- Processor: Intel core i3 1.60GHz or above
- Hard Disk: 250 GB
- RAM: 4.00 GB
- Input: Keyboard and Mouse
- Output Device: High Resolution Monitor

C. Functional Requirements

- **Data pre-processing:** The purpose of pre-processing is to check for missing values in the dataset. If any such values are found, It is replaced by mode of the corresponding values.
- **Feature Extraction:** All 32 features are input to the feature selector. This module selects a subset from the actual feature set. This process is usually done to improve accuracy and reduce the training time when the number of feature is very large.
- **Hyper tuning module:** It is here that the values of the parameters of the classifier are changed in order to increase the performance of the classifier. The parameters can be varied and the one which gives the better accuracy is selected as the model.
- **Results:** Confusion Matrix, Log Loss.

D. Non Functional Requirements

- **PERFORMANCE REQUIREMENT**, low test log loss rate has been successfully achieved using XGBoost Classifier with Hyper Parameters for both .bytes and image files individually, and as well as after merging features of .bytes and image files.

- **SOFTWARE QUALITY REQUIREMENT**, maximum possible accuracy has been achieved using XGBoost Classifier with hyperparameters using Random Search with log loss of 0.385, XGBoost Classifier with log loss of 0.0427, and Random Forest Classifier with log loss of 0.4192. It generated the confusion matrix. It used minimal resources for training the dataset as well as obtaining the results. The module is reliable and can be used to classify most of the malware in the validation set.

VI. DATA PREPROCESSING

The dataset [55] comprised a total of 5836 images (Table 1) segmented into two main parts, a training set and a test set. Both bacterial and viral pneumonia were considered as a single category, pneumonia infected. The dataset used in this study did not include any case of viral and bacterial co-infection. All chest X-ray images were taken during the routine clinical care of the patients. Two expert

physicians then graded the diagnoses for the images before being cleared for training the AI system. The evaluation set was also checked by a third expert to account for any grading errors. The proportion of data assigned to training and testing was highly imbalanced. Therefore, the dataset was shuffled and arranged into training and test sets only. Finally, there were 5136 images in the training set and 700 images in the test set. Eleven-point-nine-five percent of the complete dataset was used as the testing dataset. Figure 2 shows two chest X-ray images, one of a healthy person and the other of a person suffering from pneumonia.

In our Project, The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of paediatric patients of one to five years old from Guangzhou Women and Children's Medical Centre, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care.

For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.

Category	Training Set	Test Set
Normal (Healthy)	1283	300
Pneumonia (Viral + Bacteria)	3873	400
Total	5156	700
Percentage	88.05%	11.95%

TABLE 1: *DESCRIPTION OF THE EXPERIMENTAL DATASET.*

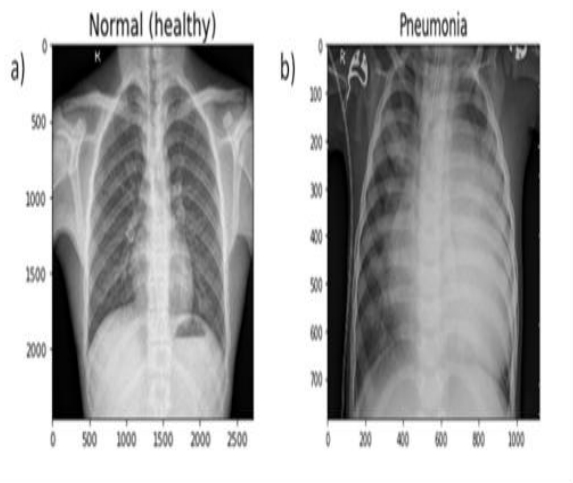

Fig No 2: *CHEST XRAY OF (A) A HEALTHY PERSON AND (B) A PERSON SUFFERING FROM PNEUMONIA.*

The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal specification in the image. Bacterial pneumonia (middle) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia (right) manifests with a more diffuse ‘‘interstitial’’ pattern in both lungs.

VII. BACKGROUND OF DEEP LEARNING METHODS

7.1: CONVOLUTION NEURAL NETWORK

LeCun et al. [52] first used CNN, in 1989, for handwritten zip code recognition. This is a type of feed-forward network. The main advantage of CNN compared to its predecessors is that it is capable of detecting the relevant features without any human supervision. A series of convolution and pooling operations is performed on the input image, which is followed by a single or multiple fully connected layers, as shown in Figure 1. The output layer depends on the operations being performed. For multiclass classification, the output layer is a softmax layer. The main disadvantage with deeper CNNs is vanishing gradients, which can be solved by using residual networks introduced in the following section.

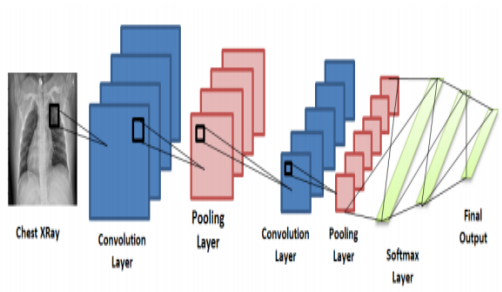

FIG No. 3: CNN CONSISTING OF CONVOLUTION AND POOLING LAYERS AND FULLY CONNECTED SOFTMAX LAYERS AT THE END TO GIVE THE FINAL PREDICTION.

7.2: TRANSFORM LEARNING

In transfer learning, a model that is trained for a particular task is employed as the starting point for solving another task. Therefore, in transfer learning, pre-trained models are used as the starting point for some specific tasks, instead of going through the long process of training with randomly initialized weights. Hence, it helps with saving the substantial computer resources needed to develop neural network models to solve these problems.

Pan and Yang [53] used domain, task, and marginal probabilities to propose a framework for better understanding the transfer learning. The domain D was defined as a two-element tuple consisting of the feature space, χ, with a marginal probability, P(X), where X is a sample data point. Hence, mathematically, domain D can be defined as,

$$D = \{\mathbf{x}, \mathbf{p}(\mathbf{x})\}$$

Here, χ is the space of all term vectors, xi is the ith term vector corresponding to some documents, and X is a particular learning sample (X = x1, • • •, xn, $\in \chi$). For a given domain D, the given task T is defined as:

$$T=\{\gamma,P(Y\gamma|x)\} = \gamma\{\gamma,\eta\}, Y=\{y1,\ldots\ldots,yn\}, yi\in\gamma$$

Where γ is the label space. η is a predictive function learned from the feature vector/label pairs (xi , yi), where xi $\in \chi$ and yi $\in \gamma$.

$$\eta(xi) = yi$$

Here, η predicts a label for each feature vector

Due to the lack of a sufficient dataset, training a deep learning based model for medical diagnosis related problems is computationally expensive, and the results achieved are also not up to the mark. Hence, pre-trained deep learning models, which were previously trained on ImageNet [54] dataset, were used in this paper. Further, all these pre-trained models were fine-tuned for pneumonia classification. All the layers of the architectures used were trainable. Further details, related to fine-tuning, and are discussed in Section 5.2.

7.3: PRE TRAINED NEURAL NETWORK

Five state-of-the-art deep learning networks, ResNet18, DenseNet121, InceptionV3, Xception, and MobileNetV2, were used in this study. They are briefly discussed in Appendix A at the end of the paper.

7.4: PERFORMANCE METRICS FOR CLASSIFICATION

All the models were tested on the test dataset after the completion of the training phase. Their performance was validated using the accuracy, recall, precision, F1, and area under the curve (AUC) score. All the performance metrics used in this paper are discussed below.

In the below-mentioned definitions and equations, while classifying healthy and pneumonia patients, true positive (TP) denotes the number of pneumonia images identified as pneumonia, true negative (TN) denotes the number of normal images identified as normal (healthy), false positive (FP) denotes the number of normal images incorrectly identified as pneumonia images, and false negative (FN) denotes the number of pneumonia images incorrectly identified as normal.

Accuracy: It tells us how close the measured value is to a known value.

$$Accuracy = (TP + FN)/ (TP + TF + FP + FN)$$

Precision: It tells about how accurate the model is in terms of those which were predicted positive

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

Recall: It calculates the number of actual positives the model was able to capture after labeling it as positive (true positive).

$$Recall = TP /(TP + FN)$$

F1: It gives a balance between precision and recall.

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

AUC Score and ROC Curve: ROC (receiver operating characteristics) is a probability curve, and AUC (area under curve) represents the degree of reparability. The ROC curve is the plot of sensitivity (true positive rate) against specificity (false positive rate).

VIII. IMPLEMENTATION

An optimum solution for the detection of pneumonia from chest X-rays is proposed in this paper. Data augmentation was used to address the problem of the limited dataset, and then, state-of-the-art deep learning models, as discussed in Section 3, were fine-tuned for pneumonia classification. Then, predictions from these models were combined, using a weighted classifier (discussed afterward in this section), to compute the final prediction. The complete block diagram of the proposed methodology can be seen in Figure 3.

After merging the features of .bytes and .asm file, malware files were reclassified using Random Forest (RF) classifier, XGBoost classifier, and XGBoost classifier with hyper parameter using Random Search, and we were able to achieve the Log loss of 0.04192, 0.0427 and 0.0385 respectively.

The settings utilized in image augmentation are shown below in Table 2. The images after performing various augmentation techniques are shown below (Figure 4). Only one of these techniques was used to generate the augmented image.

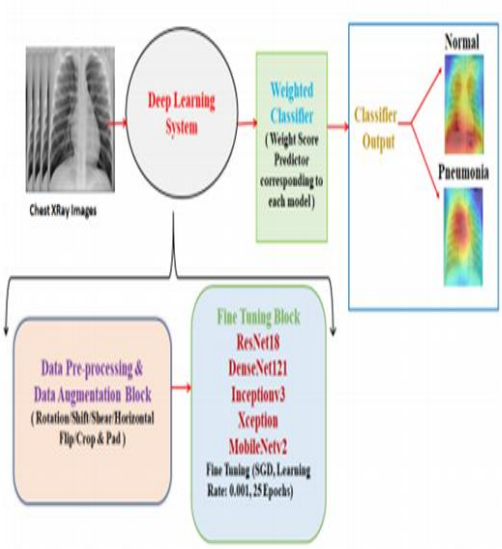

FIG NO.4: BLOCK DIAGRAM OF THE PROPOSED METHODOLOGY (DEEP LEARNING SYSTEM CONSISTS OF THE DATA PRE-PROCESSING AND DATA AUGMENTATION BLOCK AND THE FINE-TUNING BLOCK; THE WEIGHTED CLASSIFIER GIVES THE FINAL PREDICTION).

8.1: DATA PREPROCESSING AND AUGMENTATION

Each image had to be preprocessed according to the deep neural network used. There were two important steps involved: resizing and normalization. Different neural networks require images of different sizes according to their defined architecture. ResNet18, DenseNet121, and MobileNetV2 expect images of size 224 × 224, while InceptionV3 and Xception require images of size 229 × 229. All the images were also normalized according to the respective architectures.

Adequate training of a neural net requires big data. With less data availability, parameters are undermined, and learned networks generalize poorly. Data augmentation solves this problem by utilizing existing data more efficiently. It aids in increasing the size of the existing training dataset and helps the model not to over fit this dataset. In this case, there were a total of 1283 images of the normal (healthy) case and 3873 images of the pneumonia case in the training dataset. Out of these, four-hundred images were reserved for optimizing the weighted classifier. This dataset was highly imbalanced. There were already enough images in the pneumonia case. Therefore, each image of only the normal (healthy) case was augmented twice. Finally, after augmentation, there were 3399 healthy chest X-ray images and 3623 pneumonia chest X-ray images.

Technique	Setting
Rotation	45
Vertical Shift	0.2
Horizontal Shift	0.15
Shear	16
Crop and Pad	0.25

TABLE NO. 2: AUGMENTATION TECHNIQUES USED IN THE PROPOSED METHODOLOGY.

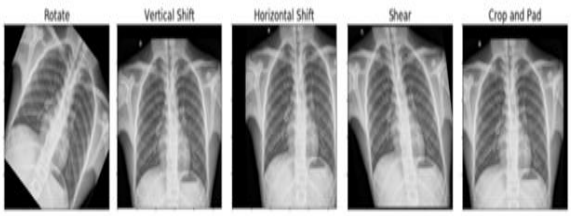

FIG NO.5: RESULTANT IMAGE AFTER PERFORMING THE AUGMENTATION TECHNIQUE.

8.2: FINE-TUNING THE ARCHITECTURES

All the architecture details used in this paper are discussed in Appendix A. Raw chest X-ray images, after being pre-processed and normalized, were used to train the network. Then, data augmentation techniques were used to process the dataset more efficiently. All the layers of the networks used were trainable, and these layers extracted the features from the images. Some parameters must be set to train the network. An

interesting paper from UC Berkeley [56] came out, and according to it, stochastic gradient descent (SGD) had better generalization than adaptive optimizers. Therefore, SGD as the optimizer was used, and the model was trained for 25 epochs. The learning rate, the momentum, and the weight decay were set to 0.001, 0.9, and 0.0001, respectively (Table 3). These configurations were to make sure that the networks were fine-tuned for pneumonia diagnosis.

Architecture	Image Size	Epochs	Optimizer	Learning Rate	Momentum	Weight Decay
ResNet18	224 × 224					
DenseNet121	224 × 224					
InceptionV3	229 × 229	25	Stochastic Gradient Descent	0.001	0.9	0.0001
Xception	229 × 229					
MobileNetV2	224 × 224					

TABLE NO.3: HYPER-PARAMETERS USED WHILE FINE-TUNING THE DEEP LEARNING MODELS

8.3: WEIGHTED CLASSIFIER

In this module of the proposed methodology, a weight (Wk) corresponding to each model was estimated. Wk can be defined as the belief in the k th model, with k being equal to 5 as 5 pre-trained models were used in this paper. Wk has values between 0 and 1, and the sum of all weights is 1 (Equation (9)). Each model, after it was fine-tuned, returned the probabilities for each class label, i.e., 2 classes in the form of a matrix (Pk). A weighted sum of all these predictions arrays was calculated (Equation (8)).

$$P1W1+P2W2+P3W3+\bullet\bullet\bullet+PkWk=Pf$$
$$W1+W2+W3+\bullet\bullet\bullet+Wk=1$$
$$Loss=-1\,N\,n\sum i=1\,y\times l(p)+(1-y)\times log(1-p)$$

Pk is the prediction matrix, with shape: number of optimization images * class labels (400*2), corresponding to each architecture. In Equation (8), the contribution of each model is weighted by a coefficient (Wk), which indicates the trust in the model. First, we obtained the Pk for every model for an unseen image set (400 images). Then, Equation (8) was optimized such that the classification error was minimized and Equation (9) was also satisfied. We used differential evolution [57] for global optimization of Equation (8). Differential evolution is a stochastic global search algorithm. It optimized Equation (8) by iteratively refining a candidate solution with regard to Equation (9). Hence, optimizing Equation (8) would provide the Wk values corresponding to each model. The value of Wk for the k th model depended on the respective models' performance on the test dataset. The maximum iterations for differential evolution algorithms were kept to be 1000. With the help of Pf, the prediction of a class label could be computed. Classification loss corresponding to this Pf was reduced while optimizing Equation (8). Log loss (Equation (10), also known as logistic loss or cross-entropy loss, was used as the loss function. In Equation (10), N denotes the size of the image set (400) and p denotes the probability that the given image is pneumonia infected. Figure 5 shows the process followed to find the optimal weight corresponding to each model. Figure 6 shows the weighted classifier used in the proposed methodology.

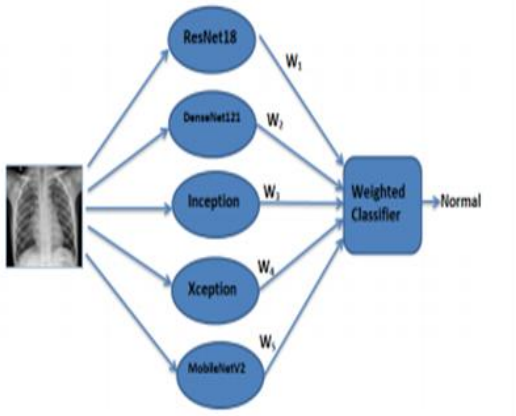

FIG NO.6: WEIGHTED CLASSIFIER MODULE USED IN THIS PAPER (WEIGHTED PREDICTIONS FROM ALL THE MODELS ARE PASSED TO THE WEIGHTED CLASSIFIER, WHICH GIVES THE FINAL WEIGHTED PREDICTION).

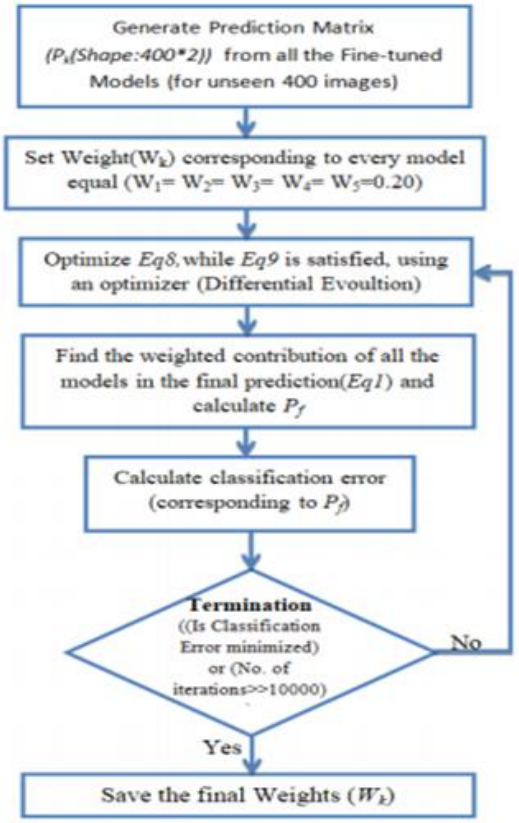

FIG NO. 7: PROCEDURE FOLLOWED WHILE CALCULATING THE OPTIMAL WEIGHT CORRESPONDING TO EVERY MODEL.

8.4: CLASS ACTIVATION MAPS

Class activation maps (CAMs) [58] can help in demystifying the results of deep learning models. Traditionally, deep learning based methods are considered to be a black-box approach. For clinical decision making, it is necessary that the results of the deep learning model can be interpreted. CAMs can help in identifying the parts of the image on which the model was focusing while making the final prediction and hence can provide insights into the working of the model. Such an analysis can further help in hyperparameters tuning and gain understanding of the reason behind the failure of the model. For obtaining the class activation map, the network needed to be trained with the global average pooling (GAP) layer. After the GAP layer, a fully connected network was maintained, which was followed by the softmax layer, providing the class, such as pneumonia, as shown in Figure 7. CAMS class activation maps were generated for both bacterial and viral pneumonia for all the fine-tuned model and are discussed in detail in the results section.

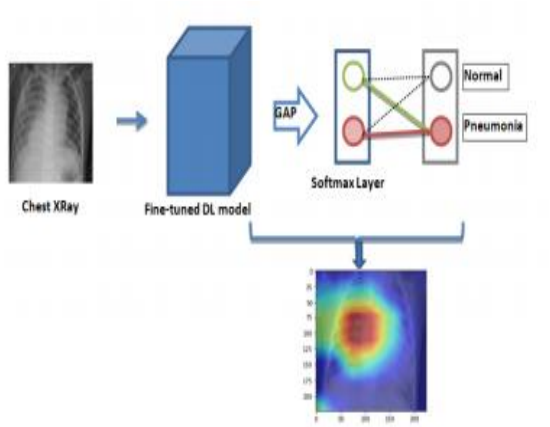

FIG NO.8 : GENERATION OF THE CLASS ACTIVATION MAP FOR FINE-TUNED DEEP LEARNING MODELS. THE LAYERS OF DEEP LEARNING MODELS ARE FOLLOWED BY THE GLOBAL AVERAGE POOLING LAYER (GAP) (⇒) AND THE SOFTMAX LAYER TO GIVE THE FINAL PREDICTION. FEATURES THAT ARE USED FOR PNEUMONIA DETECTION GET HIGHLIGHTED IN THE CLASS ACTIVATION MAP.

IX. FINAL RESULT

In this section, the experiments and evaluation techniques used in the paper to test the efficiency of the proposed model are presented. The chest X-ray image dataset, proposed in [55], was used. The Keras open-source deep learning framework with TensorFlow as the backend was used, first to load the pre-trained architectures on the ImageNet Dataset [54] and then fine-tune them for the task at hand. All the computation work was done on a Standard PC with8 GB RAM, NVIDIA GeForce GTX 1060 6 GB GPU, and Intel i7, seventh-generation processor.

9.1: RESULT IN TERMS OF TESTING ACCURACY AND TESTING LOSS

To test and evaluate the performance of the proposed network, each experiment was conducted five times. Parameters and hyperparameters were tuned during the training. Figure 8 shows the training accuracy and training loss curves obtained while training the models for 25 epochs. The training accuracy for all the models exceeded 99%, and the training loss for all the models was below 0.03. Except for Xception, all the other models had similar training accuracy and training loss curves. Table 4 summarizes the testing accuracy and testing loss for different networks and the final weighted classifier. DenseNet121 was able to attain the maximum testing accuracy and the minimum testing loss. Initially, all the weights of the weighted classifier were kept equal (W1 = W2 • • • W5 = 0.20). Hence, every model

contributed equally towards the final prediction. A test accuracy of 97.45 and a loss of 0.087 was obtained. Then, the optimum weights were estimated for every model. The value of these estimated weights is shown in Table 5. With these weights, the final weighted classifier was able to achieve a testing accuracy of 98.43, and the testing loss was 0.062.

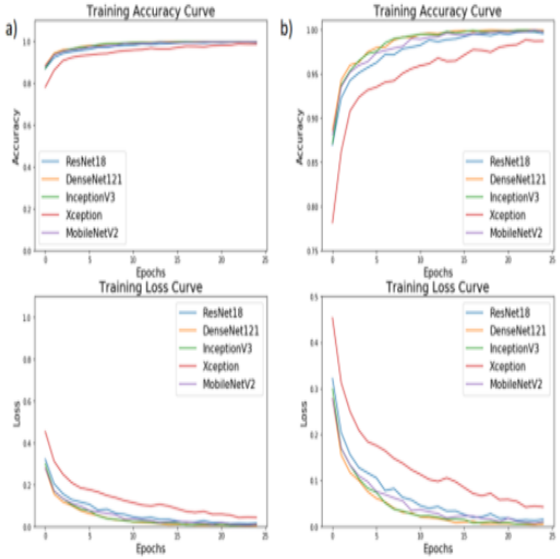

FIG NO. 9 : (A) TRAINING ACCURACY AND TRAINING LOSS CURVES FOR DIFFERENT ARCHITECTURES OVER THE TRAINING DATASET WHILE THE MODELS WERE TRAINED FOR 25 EPOCHS, (B) ZOOM-IN VERSION OF (A).

Architecture	Testing Accuracy	Testing Loss
ResNet18	97.29	0.096
DenseNet121	98.00	0.064
Inception	97.00	0.098
Xception	96.57	0.101
MobileNetV2	96.71	0.096
Weighted Classifier (With Equal Weights)	97.45	0.087
Weighted Classifier (With Optimized Weights)	98.43	0.062

TABLE NO.4: FINAL TESTING ACCURACY AND TESTING LOSS ACHIEVED BY ALL THE ARCHITECTURES AND THE WEIGHTED CLASSIFIER

Architecture	Weight
ResNet18 (W1)	0.25
DenseNet121 (W2)	0.30
Inception (W3)	0.18
Xception (W4)	0.08
MobileNetV2 (W5)	0.19

TABLE NO.5: WEIGHT VALUE (BELIEF OR TRUST VALUE) CORRESPONDING TO EVERY ARCHITECTURE

In Table 4, it can be seen that when equal weights were assigned to every model, the testing accuracy of the weighted classifier was less than that of DenseNet121. This could be attributed to the fact that even the models with less testing accuracy were assigned the same weights as that to the models with higher testing accuracy. Finally, when optimum weights were calculated, the testing accuracy of the weighted classifier showed an improvement of 0.98%. Table 5 shows that the weight assigned to every model depended on its performance on the test dataset. Hence, it could be said that the weight assigned to a model represented the belief or trust in that model. The maximum weight was assigned to DenseNet121, which was followed by ResNet18.

All the test images were pre-processed similarly as the training images and hence had the same size as required by the respective architecture. The test images were of size 224 × 224 for ResNet18, DenseNet121, and MobileNetV2, while for InceptionV3 and Xception, they were of size 229 × 229. The testing was also done on the same system on which training was done. The average inference time for all the models was 0.045 s (while the GPU was used), and for the weighted classifier, it was 0.203 s.

9.2: PERFORMANCE ANALYSIS

To further test the robustness of the proposed methodology, the accuracy, recall, precision score, F1 score, and AUC score for all the models and the proposed weighted classifier were calculated. To calculate the mentioned scores, confusion matrices for all the architectures were obtained (Figure 9). With the help of the confusion matrix, the number of true positives, true negatives, false positives, and false negatives could be calculated, which further helped in checking the efficacy of the model.

As the recall was increased, the precision decreased, and vice versa. In medical applications, all the patients who had the disease needed to be identified, and hence, the recall could be maximized.

A low recall could be accepted if the cost of a follow-up medical examination was not high. Hence, the F1 score could be used to find the optimal blend of precision and recall.

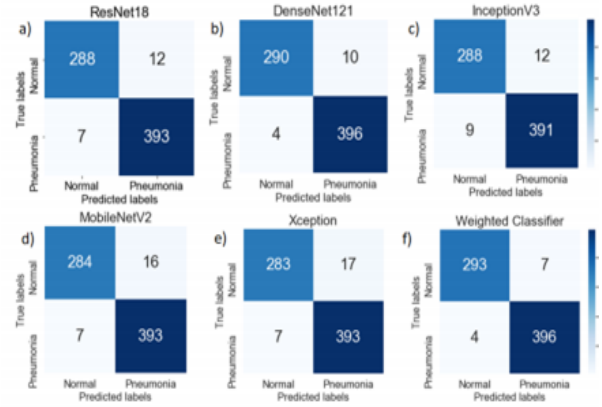

FIG NO. 10 : CONFUSION MATRIX FOR (A) RESNET18, (B) DENSENET121, (C) INCEPTIONV3, (D) MOBILENETV2, (E) XCEPTION, AND (F) WEIGHTED CLASSIFIER ARCHITECTURES AND THE WEIGHTED CLASSIFIER OVER THE TESTING DATASET. FALSE POSITIVES WERE GREATER THAN THE FALSE NEGATIVES FOR ALL THE MODELS.

In the plotted confusion matrices (Figure 9), it can be seen that the proposed weighted classifier outperformed all the individual models. The generic image features, learned by the deep learning models from ImageNet, served as a good initialization of the weights. The misclassification error for normal (healthy) images as pneumonia images was greater compared to pneumonia images as healthy images. This might be because the number of chest X-ray images of the normal (healthy) case was significantly lower compared to the pneumonia cases.

Figure 10 shows the ROC curves for different architectures and the proposed classifier. The maximum AUC score (99.76) was achieved by the proposed classifier. All the models had a similar AUC/ROC curve. All the results are tabulated in Table 6. After analyzing the results, it can be said the weighted classifier gave the best results with an AUC score of 99.75, F1 score of 98.63, and test accuracy of 98.43. Hence, the proposed weighted classifier was able to combine the predictions from all the individual architectures in an optimum manner. The differences in the performance of other models were not significant. This might be because all the models used in this paper were deep learning based and were fine-tuned on the same insufficient dataset.

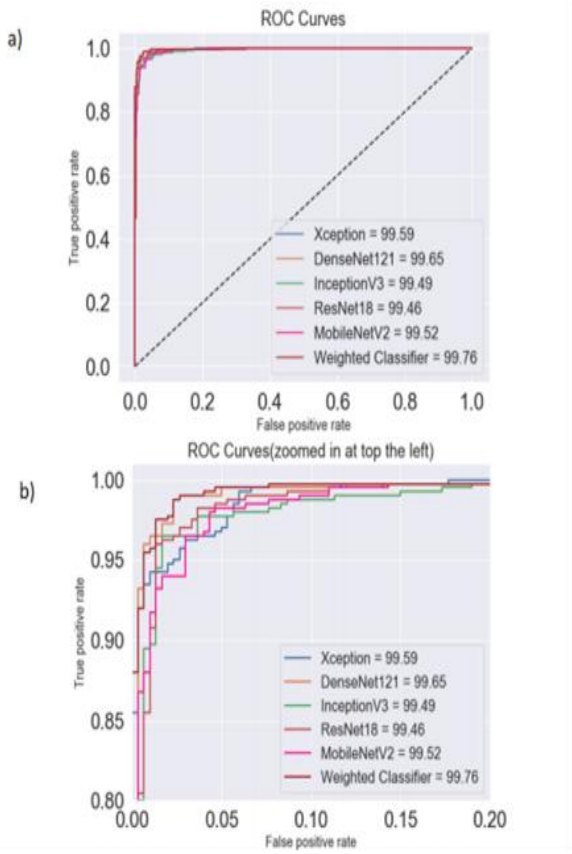

FIG NO.11: . COMPARISON BETWEEN (A) THE AUC (AREA UNDER THE CURVE) AND (B) ROC (RECEIVER OPERATING CHARACTERISTICS) CURVES FOR DIFFERENT ARCHITECTURES AND THE WEIGHTED CLASSIFIER FOR THE TESTING DATASET. THE WEIGHTED CLASSIFIER (BROWN LINE), FOLLOWED BY DENSENET121 (YELLOW LINE), HAD THE HIGHEST AUC

Architecture	Accuracy	Precision	Recall	F1 Score	AUC Score
ResNet18	97.29	97.03	98.25	97.63	99.46
DenseNet121	98.00	97.53	99.00	98.26	99.65
InceptionV3	97.00	97.02	97.75	97.39	99.49
Xception	96.57	95.85	98.25	97.03	99.59
MobileNetV2	96.71	96.08	98.25	97.15	99.52
Weighted Classifier	98.43	98.26	99.00	98.63	99.76

TABLE NO.6: ACCURACY, PRECISION, RECALL, F1 SCORE, AND AUC SCORE CORRESPONDING TO DIFFERENT ARCHITECTURES

9.3: EXPLANATION OF THE RESULTS USING HEAT MAPS

The activation maps were plotted for every individual network. These activation maps helped in localizing areas in the image most indicative of pneumonia. The activation maps were obtained for

the last convolutional layer of each network. In the case of bacterial pneumonia (Figure 11), all the networks detected the abnormal lung to predict the presence of pneumonia correctly. Viral pneumonia manifested with a more diffuse "interstitial" pattern in both lungs, which was detected by all the fine-tuned architectures [59] (Figure 12).

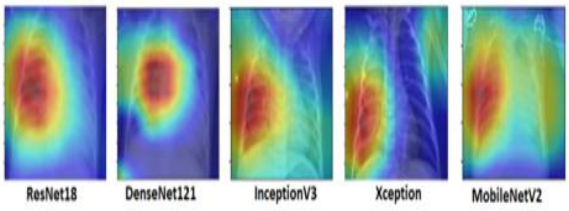

FIG NO.12: ACTIVATION MAPS FOR CHEST X-RAYS HAVING BACTERIAL PNEUMONIA CORRESPONDING TO DIFFERENT ARCHITECTURES. ABNORMAL LUNGS, IN THE CASE OF BACTERIAL PNEUMONIA, WERE DETECTED BY THE DEEP LEARNING MODELS

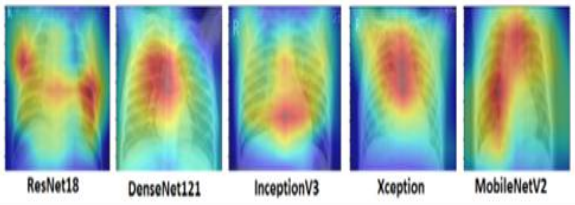

FIG NO.13: ACTIVATION MAPS FOR CHEST X-RAYS HAVING VIRAL PNEUMONIA CORRESPONDING TO DIFFERENT ARCHITECTURES. VIRAL PNEUMONIA, WITH A MORE DIFFUSED "INTERSTITIAL" PATTERN IN BOTH LUNGS, WAS DETECTED BY THE DEEP LEARNING MODELS.

9.4: COMPARATIVE ANALYSIS OF VARIOUS EXISTING METHODS

The accuracy of various existing methods and the proposed methodology were compared. All the results mentioned in this section are reported by the authors in their respective studies. Rahib H. Abiyey et al. [36] used CNN and achieved a validation accuracy of 92.4%.

The test dataset used was smaller compared to this paper. Okeke Stephen et al. [37] achieved a validation accuracy of 93.73% with their own CNN model. No other metric was published in either of these works. Cohen et al. [38] used a model based on DenseNet-121. They reported an AUC score of 98.4%. Unfortunately, the other metrics were not reported in the paper. Rajaraman et al. [39] used customized CNNs to detect pneumonia and reported a test accuracy of 96.2%. M.Togacar et al. [60] combined features from different deep learning models for pneumonia classification and achieved an accuracy of 96.84%. Vikash et al. [51] combined the outputs of different neural networks and reached the final prediction using majority voting. They achieved an AUC score of 99.34. Saraiva et al. [44], Ayan et al. [45], and Rahman et al. [46] used deep learning based methods and achieved an accuracy of 94.4%, 84.5%, and 98.0%, respectively. In all of these papers, the dataset used was of a similar size. All the studies other than Rahib H.Abiyey et al. [36] used image augmentation techniques. All the above-discussed results are summarized in Table 7.

X. FUTURE WORK

The high test accuracy (98.43) and AUC score (99.76) showed that the proposed method could be used as a supplement in clinical decision making. It can only aid the radiologists in the decision making process; the final decision has to be made by an expert. The proposed weighted classifier, with optimum weights, showed an improvement of 0.98%, in terms of the testing accuracy, over the case in which equal weights were assigned to every model. The false positives were greater than the false negatives, and hence, the classification error of pneumonia suffering patients as healthy was comparatively lesser, which is ideally required in medical diagnosis. Further, the activation maps plotted in this paper showed that the deep learning based models used were able to identify pneumonia affected regions in the chest X-rays. When compared to DenseNet121, the proposed weighted classifier showed an improvement of 0.43% in terms of testing accuracy, which in the real world on a large test dataset would be a significant number.

One of the limitations of this approach was the scarcity of available data. Usually, deep learning models are trained over thousands of images. Training deep neural networks with limited data might lead to overfitting and restricts the models' generalization ability. Unlike large datasets like ImageNet, the variability in the chest X-ray data was several orders of magnitude smaller. The performance of the proposed methodology would only increase with the availability of more data. Another limitation was that the results of the deep learning models could not be properly explained. A deep understanding of the radiological features visible in chest X-rays is required for the diagnosis of the disease from the X-rays. The proper explanation of the final prediction of the model is also required, and this is one of the drawbacks of the deep learning based models. To this end, the activation maps were plotted, but further work is required. In the future, with better annotated datasets available, deep learning based methods might be able to solve this problem.

XI. CONCLUSION

Pneumonia constitutes a significant cause of morbidity and mortality. It accounts for a considerable number of adult hospital admissions, and a significant number of those patients ultimately die (with a mortality rate of 24.8% for patients over 75 years) [61]. According to the WHO, pneumonia can be prevented with a simple intervention and early diagnosis and treatment [4]. Nevertheless, the majority of the global population lacks access to radiology diagnostics [62]. Even when there is the availability of imaging equipment, there is a shortage of experts who can examine X-rays.

Through this paper, the automatic detection of pneumonia in chest X-ray images using deep transfer learning techniques was proposed. The deep networks, which were used in our methodology, had more complex structures, but fewer parameters and, hence, required less computation power, but achieved higher accuracy. Transfer learning and data augmentation were used to solve the problem of over fitting, which is seen when there is insufficient training data, as in the case of medical image processing. Further, to combine different architectures efficiently, a weighted classifier was proposed. The experiments were performed, and the different scores obtained, such as the accuracy, recall, precision, and AUC score, proved the robustness of the model. The proposed model was able to achieve an accuracy of 98.857%, and further, a high F1 score of 99.002 and AUC score of 99.809 affirmed the efficacy of the proposed model. Though many methods have been developed to work on this dataset, the proposed methodology achieved better results. In the future, it would be interesting to see approaches in which the weights corresponding to different models can be estimated more efficiently and a model that takes into account the patient's history while making predictions.

REFERENCES

[1]. Johns Hopkins Medicine. Pneumonia. Available online: https://www.hopkinsmedicine.org/health/ conditions-and-diseases/pneumonia (accessed on 31 December 2019).

[2]. Johnson, S.; Wells, D. Healthline. Viral Pneumonia: Symptoms, Risk Factors, and More. Available online: https://www.healthline.com/health/viral-pneumonia (accessed on 31 December 2019).

[3]. Healthcare, University of Utah. 2016. Pneumonia Makes List for Top 10 Causes of Death. Available online: https://healthcare.utah.edu/the-scope/shows.php?shows=0_riw4wti7 (accessed on 31 December 2019).

[4]. https://towardsdatascience.com/malware-classification-using-machine-learning-7c648fb1da79

[5]. WHO. 2011. Pneumonia is the Leading Cause of Death in Children. Available online: https://www.who. int/maternal_child_adolescent/news_events/news/2011/pneumonia/en (accessed on 31 December 2019).

[6]. Rudan, I.; Tomaskovic, L.; Boschi-Pinto, C.; Campbell, H. Global estimate of the incidence of clinical pneumonia among children under five years of age. Bull. World Health Organ. 2004, 82, 895–903. [PubMed]

[7]. Pneumonia. Available online: https://www.radiologyinfo.org/en/info.cfm?pg=pneumonia (accessed on 31 December 2019).

[8]. World Health Organization. Standardization of Interpretation of Chest Radiographs for the Diagnosis of Pneumonia in Children; Technical Report; World Health Organization: Geneva, Switzerland, 2001.

[9]. Cherian, T.; Mulholland, E.K.; Carlin, J.B.; Ostensen, H.; Amin, R.; Campo, M.D.; Greenberg, D.; Lagos, R.; Lucero, M.; Madhi, S.A.; et al. Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. Bull. World Health Organ. 2005, 83, 353–359.

[10]. Franquet, T. Imaging of pneumonia: Trends and algorithms. Eur. Respir. J. 2001, 18, 196–208. [CrossRef]

[11]. Tahir, A.M.; Chowdhury, M.E.; Khandakar, A.; Al-Hamouz, S.; Abdalla, M.; Awadallah, S.; Reaz, M.B.I.; Al-Emadi, N. A systematic approach to the design and characterization of a smart insole for detecting vertical ground reaction force (vGRF) in gait analysis. Sensors 2020, 20, 957. [CrossRef] [PubMed]

[12]. Chowdhury, M.E.; Alzoubi, K.; Khandakar, A.; Khallifa, R.; Abouhasera, R.; Koubaa, S.; Ahmed, R.; Hasan, A. Wearable real-time heart attack detection and warning system to reduce road accidents. Sensors 2019, 19, 2780. [CrossRef]

[13]. Chowdhury, M.E.; Khandakar, A.; Alzoubi, K.; Mansoor, S.; M Tahir, A.; Reaz, M.B.I.; Al-Emadi, N. Real-Time Smart-Digital Stethoscope System for Heart Diseases Monitoring. Sensors 2019, 19, 2781. [CrossRef]