Convolutional Neural Network Based Chest X-Ray Image Classification for Pneumonia Diagnosis

Rushi Bhatt(✉), Sudhanshusinh Yadav(✉), and Jignesh N. Sarvaiya(✉)

Electronics Engineering Department, SVNIT, Surat 395007, India
rybhatt27@gmail.com, sudhanshusingh8570@gmail.com, jns@eced.svnit.ac.in

Abstract. Pneumonia is one of the most chronic diseases, and therefore, its timely diagnosis is of utmost importance. Traditionally, clinical decisions have been considered as a gold standard for diagnosis, but it is not a practical option in all scenarios. Therefore, several methods have been explored to make the process of diagnosis faster, efficient and as accurate as clinical decisions. In this paper, we have described and proposed a Convolutional Neural Network (CNN) based deep learning technique for the classification of chest X-ray images for the diagnosis of Pneumonia. The proposed model is trained on 4099 images and tested on 1757 images resulting in an accuracy of 96.18%. The evaluation and training are conducted on *'Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification'* dataset which is one of the largest labeled datasets which is publicly available. Also, a comparison of the proposed model with various other popular models is discussed. The results indicate that our model despite having simpler architecture and without any pre-training outperforms many of the popular models on several different performance parameters.

Keywords: Pneumonia · Deep learning · CNN · Chest X-Ray · Diagnosis

1 Introduction

Chest diseases are one of the major health problems in particular pneumonia is extremely dangerous for people already suffering from other diseases, infants, and older adults. According to WHO evaluation, 450 million cases of pneumonia are registered every year that is 7% of the total world population. Moreover, nearly 4 million people die because of it. This ratio is even higher for infants and older adults [1]. That is why Pneumonia requires proper diagnosis at the initial stages for recovery treatment.

Traditionally, diagnosis of pneumonia has been done by medical specialists using chest X-rays and sophisticated radiological investigation on them. However, this approach, for appropriate analysis, would require radiologists. For instance, the laboratory diagnosis of these ailments involves the detection of pathogens such as a virus in the upper and lower respiratory by the use of microscopy techniques [1]. World Health Organization (WHO) has estimated that approximately only one-third of the world

S. Gupta and J. N. Sarvaiya (Eds.): ET2ECN 2020, CCIS 1214, pp. 254–266, 2020.
https://doi.org/10.1007/978-981-15-7219-7_22

population has access to a radiologist for the diagnosis of their disease. Therefore, to bring about a solution to these challenges, various computerized systems have been developed to analyze these X-ray images for medical diagnosis [2].

The computerized technique has been adopted because they give more precise results and are easily accessible for diagnosis. The image processing technique has a powerful ability to detect various objects together, extract deep features and classifying them [3] and therefore it is popularly the initial step achieved by convolutional layers of CNN. Traditionally in the interpretation of pneumonia, a radiologist looks for a white spot in the lung, which represents infection. This step of observing patterns, similarities and dissimilarities is achieved by the kernels of CNN and is a crucial step in classification problems.

The major problem in the medical domain is the lack of availability of large image datasets. Particularly for pneumonia detection, the task is very tedious because of the multiple and diverse nature of the disease. This leads to the development of transfer-learning based pre-trained models for medical imaging-based classification problems. The Deep Convolutional Neural Network shows the potential for highly variable tasks across many object categories [4] and therefore selecting appropriate hyperparameters is of utmost importance. However, in this paper, we have outperformed transfer-learning based models using a simpler architecture purely by precisely tuning hyper-parameters of the model.

2 Literature Review

In this work [5], General Regression Neural Network proposed for the prediction of active pulmonary TB. Input parameters are divided into three groups: demographic variables, constitutional symptoms, and radiographic findings. The model consists of three layers: input, hidden, and output layer, where the hidden layer is used to extract higher-order features, and the output layer gives the probability of active pulmonary TB. This model achieved the specificity of 69% on the validation dataset.

This paper [6], focuses on chest disease diagnosis using several neural network architectures. The analysis was done for six different chest diseases, out of which the best accuracy was obtained for Pneumonia diagnosis using a Multi-Layer NN (MLNN) model. The model achieves an accuracy of 91.67% for a single hidden layer and 90.00% for two hidden layers.

Because of inefficiency in working with high-dimensional image datasets, various deep learning models have been developed for the diagnosis of various diseases. CheXNet [7] uses a 121-layer pre-trained CNN and was extended to detect 14 diseases from ChestX-ray14 [8] dataset. The accuracy achieved for the binary classification problem of pneumonia detection is 76.80%.

Another deep learning model called ChetNet [9] was developed for the diagnosis of 14 thorax diseases. The proposed model comprises of classification and attention branch, where the classification branch implements feature extraction and attention branch calculates the correlation between class label and location of abnormalities. The final diagnosis is achieved by averaging the output of both branches. This proposed

method achieves an accuracy for pneumonia detection is 69.75%, while the average per-class accuracy encountering all the thorax diseases is 78.10%.

The availability of pre-trained models led to the development of various transfer-learning based models for diagnosis of Pneumonia. The paper [10] describes a generalized model primarily capable of performing diagnosis through OCT image analysis. It was further extended to diagnose pneumonia based on chest X-ray images. The model was pre-trained on the ImageNet dataset and then was fine-tuned for the desired application. This model achieved an accuracy of 92.8% with a sensitivity of 93.2% and a specificity of 90.1%.

Recently, various CNN based models have been developed for the detection of pneumonia from a chest X-ray image. The work [11] proposes a model consisting of feature extraction and classification. The feature extractor consists of four convolution layers, and the classifier is a simple feed-forward network. The best performing validation accuracy by this model is 93.73%.

Another CNN based binary classifier [12] uses Chest X-Ray Images for diagnosis of Pneumonia. It uses k-fold cross-validation for evaluation of the performance of the model, obtaining an average accuracy of 95.30%.

Before this paper, a lot of work has been done in the field of diagnosis of Pneumonia by using chest X-ray images. The ANN-based architectures [5, 6] fail to produce high accuracy results because of the lack of ANNs to deal with high-dimensional features from images for better generalizability. Later, various CNN based models [11, 12] have been developed to tackle the challenges of overfitting due to the huge number of training parameters resulting in problems of overfitting. Moreover, even some transfer-learning based models [10] have been utilized due to the availability of pre-trained models. In this paper, we propose a 9-layer CNN based model that is trained on the same dataset as several of the models, as mentioned earlier, achieving a 96.18% validation accuracy with minimum trainable parameters.

3 Methodology

3.1 Proposed Model

Figure 1 shows the proposed CNN architecture for the two-class problem of pneumonia diagnosis. The proposed model consists of nine layers out of which there are two 2D-Convolutional layers, two Pooling layers, one layer each for Batch Normalization, Dropout, and Flatten and then finally two Dense (Fully Connected) layers leading to a SoftMax output.

3.2 Input Preprocessing

The first step involves resizing the input image to a $64 \times 64 \times 3$-dimensional matrix. Then, the 8-bit representation of each image is normalized to a scale of 0 to 1 by dividing each pixel value by 255.

For an input image 'X', each pixel at position (i, j), the zero-centered standardization is expressed in terms of the mean '$\bar{X}$' and standard deviation 'σ_X' as shown in the Eq. (1)

$$X_{stand}(i,j) = \frac{X(i,j) - \bar{X}}{\sigma_X} \tag{1}$$

These input scaling methods remove the biases that might have been introduced due to abnormal deformities in the X-ray images.

3.3 Description of the Architecture

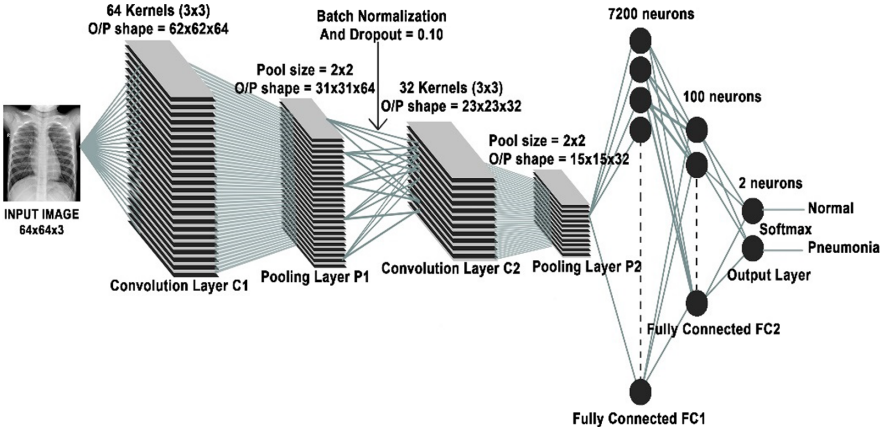


Fig. 1. Proposed CNN architecture

As shown in Fig. 1, the first convolutional layer 'C1' comprises of 64 kernels, each of size 3 × 3. Moreover, the kernel is initialized by the 'glorot uniform' distribution function, which unlike random initialization, helps in achieving the global minimum of the loss function faster with relatively less training.

The initial weight values 'W' for the j^{th} layer is dependent on the number of weights in j^{th} layer 'n_j' as well as the next layer 'n_{j+1}' according to the Uniform Distribution function 'U' as shown in Eq. (2).

$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, +\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}\right] \tag{2}$$

The layer immediately after 'C1' is the first 2D-pooling layer 'P1' implementing maxpool with a pool size of 2 × 2. Including 'P1' helps in the reduction of dimension as well as the complexity of the network and also avoids overfitting, thereby making the model generalized.

Following the 'P1' layer is the Batch Normalization layer, which normalizes the entire batch of inputs from the previous layer in a slightly modified way, as described in Eq. (1) for the input image. This layer helps in speeding up convergence and accelerating the training by reducing the internal covariate shift [13]. Moreover, a dropout layer is also added, which is set to drop 10% of the connection to the next layer to prevent overfitting [14].

The next layers consisting of a convolutional layer 'C2' with 32 kernels each of size 3×3, and following it is the pooling layer 'P2'. The output of 'P2' is flattened and then fed as input to the fully connected neural network layers and finally classifying the image into two classes: Normal and Pneumonia. The first Fully Connected layer 'FC1' has 7200 input neurons and 100 output neurons, each a 'ReLu' activation.

'ReLu' stands for the Rectified Linear Unit and is also used as activation for the convolutional layers 'C1' and 'C2'. The output of ReLu 'Φ' can be defined in terms of the input 'x' as described in the Eq. (3)

$$\Phi(\mathrm{x}) = \begin{cases} x; x > 0 \\ 0; x \leq 0 \end{cases} \tag{3}$$

The next Fully Connected layer 'FC2' has 100 inputs from 'FC1' and 2 output, one for each class having a 'SoftMax' activation. The output of SoftMax describes the probability 'Φ' for each class 'j' in terms of input 'x' as described in the Eq. (4) (Table 1).

$$\Phi(\mathrm{x_j}) = \frac{\mathrm{e}^{\mathrm{x_j}}}{\sum_{\mathrm{i}=1}^{2} \mathrm{e}^{\mathrm{x_i}}} \tag{4}$$

Table 1. Description of the proposed CNN architecture comprising of a total of 740,714 trainable parameters and 128 non-trainable parameters

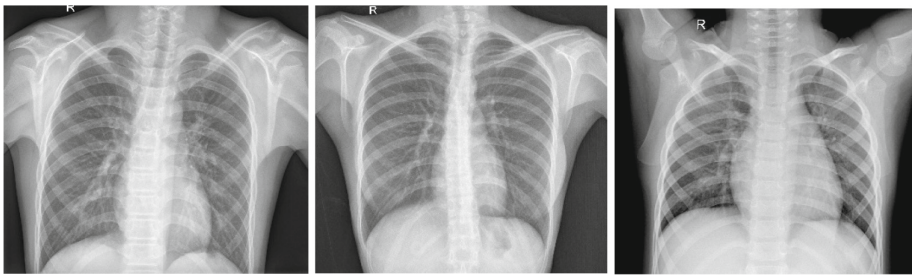
Layer (Type)	Output shape	No. of parameters
conv2d_1 (Conv2D)	(None, 62, 62, 64)	1792
max_pooling2d_1 (MaxPooling2)	(None, 31, 31, 64)	0
batch_normalization_1 (Batch Normalization)	(None, 31, 31, 64)	256
dropout_1 (Dropout)	(None, 31, 31, 64)	0
conv2d_2 (Conv2D)	(None, 29, 29, 32)	18464
max_pooling2d_2 (MaxPooling2)	(None, 15, 15, 32)	0
flatten_1 (Flatten)	(None, 7200)	0
dense_1 (Dense)	(None, 100)	720000
dense_2 (Dense)	(None, 2)	202

3.4 Dataset

The dataset [15] used for training, testing, and validating the model contains a total of 5863 chest X-ray Images (JPEG file format) belonging to two categories: Normal and

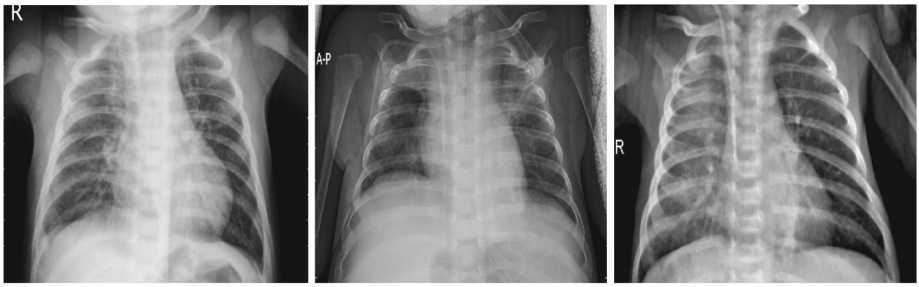
Pneumonia. Pneumonia class further has two subclasses of Pneumonia, which are Bacterial and Viral.

The labeling of these chest X-ray radiographs was carried out by two expert physicians, followed by a third expert to avoid discrepancies due to grading errors. Some of the sample images of chest X-ray from the dataset [15] have been shown in Fig. 2 and Fig. 3.



(a) (b) (c)

Fig. 2. (a), (b) and (c): Sample images without pneumonia (Normal)



(a) (b) (c)

Fig. 3. (a), (b) and (c): Sample images with pneumonia

3.5 Computational Logistics

All the simulation was performed on 'Spyder IDE' with all codebase written in 'Python3'. The primary deep learning frameworks that have been used are 'Keras 2.2.4' and 'Tensorflow 1.13.1' to build and train the convolutional neural network model. All the experiments were run on a standard PC with Nvidia GeForce 1050 Ti GPU card of 4 GB (DDR5) with a DDR4 RAM of 16 GB.

3.6 Training and Testing

The entire database [15] on which the model training, as well as model evaluation, is performed consists of 1583 Normal Images and 4273 Pneumonia Images. Out of these, the model is trained on 70% of the dataset, and the remaining 30% is reserved for testing and validation.

The proposed CNN based architecture consists of several parameters (or weights) distributed throughout the layers which are optimized and used for constructing the classification model. Out of all the layers shown in Fig. 1, the layer 'C1', Batch Normalization, 'C2', 'FC1', and 'FC2' contribute to the weight parameters. Out of all these layers, 'C2' contributes the most with 720,000 parameters. The entire proposed architecture has a total of 7,40,714 parameters, out of which 7,40,586 are trainable, and 128 are non-trainable parameters.

The training process uses categorical cross-entropy as the loss function. The reason for using this loss function is its robustness against noisy labels [16].

The categorical cross-entropy loss 'H' can be calculated based on the predicted categorical output by the model 'Q' and the ground-truth categorical labels 'P' for 'N' classes as shown in the Eq. (5).

$$H(P,Q) = -\sum_{i=1}^{N} P(i) \cdot \log(Q(i)) \tag{5}$$

Furthermore, a stochastic optimizer - 'Adam' is used because of its ability to work well in a sparse setting with low-resource requirements [17] with a stepsize α of 0.01 and exponential decay rate for moment estimates β_1 and β_2 of 0.9 and 0.999 respectively. The hyperparameters chosen for developing the model are summarized in Table 2.

Table 2. Hyperparameters used in training the model

Parameter name	Value
Learning rate	0.0001
Dropout rate	0.1
Optimizer used	Adam
Convolutional layer activation	ReLu
'FC1' Activation	ReLu
'FC2' Activation	SoftMax
Batch size	32

4 Results

4.1 Confusion Matrix

The confusion matrix is an 'N' $\times$ 'N' size matrix, where 'N' is the total number of classes, representing the number of correct as well as misclassifications. In our binary

classifier, the confusion matrix is a 2 × 2 matrix consisting of each of the following parameters as defined below:

- True Positive (TP) is defined as the number of true pneumonia X-ray images correctly predicted by the classifier.
- True Negative (TN) is defined as the number of true normal X-ray images correctly predicted by the classifier.
- False Positive (FP) is defined as a number of true normal X-ray images wrongly predicted by the classifier.
- False Negative (FN) is defined as the number of true pneumonia X-ray images wrongly predicted by the classifier.

As shown in Fig. 4, the TP are 1225 contributing to the probability of 0.98 of correctly identifying Pneumonia, and the TN is 461 contributing to a probability of 0.91 to classify normal samples correctly.

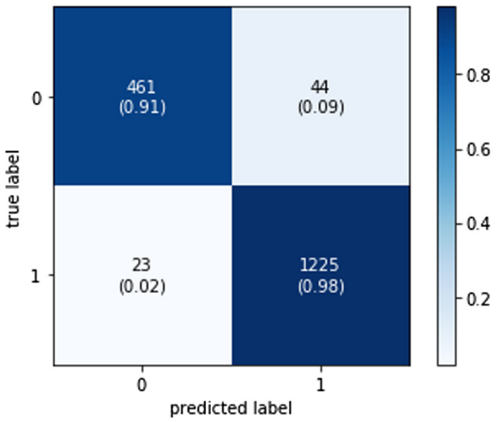


Fig. 4. Confusion Matrix evaluation on the testing dataset (Class 0 represents Normal X-ray Chest Images, and Class 1 represents Pneumonia X-ray Chest Images)

4.2 Fundamental Performance Parameters

Some of the fundamental parameters which are used for performance evaluation of a classifier are Sensitivity, Specificity, Precision, and F1 score.

Sensitivity is the probability of images that were classified as pneumonia out of the total number of pneumonia image samples and also known as the Recall parameter or TPR (True Positive Rate). Specificity is the ratio of the correctly classified normal images out of the total number of normal samples. Precision is defined as the ratio of the total number of correctly classified pneumonia samples out of total samples that were classified as pneumonia [18]. The mathematical equations of sensitivity, specificity, and precision are as shown in Eq. (6), Eq. (7) and Eq. (8) respectively.

$$\text{Re}call = TPR = Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

F1 Score is a metric that takes into account recall as well as the precision and is the weighted average of precision and recall. F1 score gives more information than the ROC curve for binary classifiers [19]. The Performance Summary for the Binary Classifier is given in Table 3.

$$F1\ Score = 2\left(\frac{Recall \times Precision}{Recall + Precision}\right) \tag{9}$$

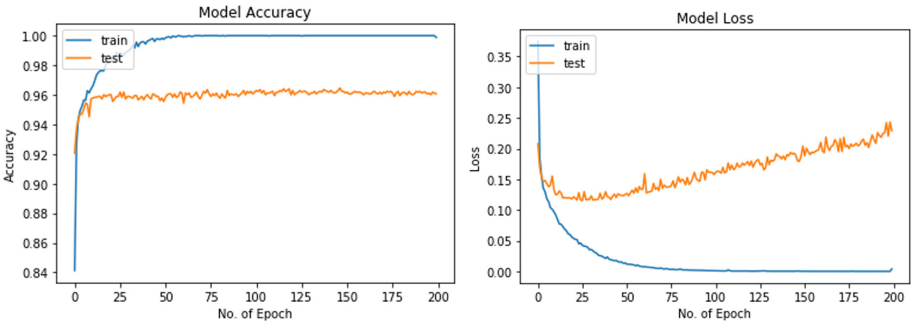
Table 3. Performance Summary for the proposed binary classifier

	Precision	Recall	F1 score	Validation samples
Normal class	0.95	0.91	0.93	505
Pneumonia class	0.97	0.98	0.97	1248

4.3 Validation Accuracy and Loss

The model evaluation was done by testing it on 30% of the total dataset on which it had never been trained. In addition, the model accuracy strongly affected by the values of hyperparameters, which has been discussed in Sect. 5.1.

The results obtained are training accuracy = 1.00, testing accuracy = 0.9618 and the plots of accuracies and losses after each epoch are as shown in Fig. 5 (Table 4).



(a) (b)

Fig. 5. Training and Testing accuracy and loss plots

Table 4. Performance comparison between [10] and our proposed model on the same dataset

Performance metric (%)	*D. S. Kerman etc.* [10]	*Proposed model*
Validation accuracy	92.8	96.18
Sensitivity	93.2	98.16
Specificity	90.1	91.29
Area under ROC	96.8	98.91

4.4 ROC Curve and AUC

ROC (Receiver Operating Characteristics) and AUC (Area Under the Curve) are some of the most crucial performance measurement parameters for classification models. ROC represents the classifier performance across the entire distribution of classes [20] and the AUC the area under ROC, which is the measure of separability or the model's capability to distinguish between classes. Higher the AUC, the better the model can identify the categories and therefore it is an essential parameter for characterizing the strength and weakness of diagnostic tests [21].

The area under the ROC curve or AUC, achieved by our model, is 0.9891. Figure 6 shows the ROC curve plotted between True Positive Rate and False Positive Rate on the testing data for our model.

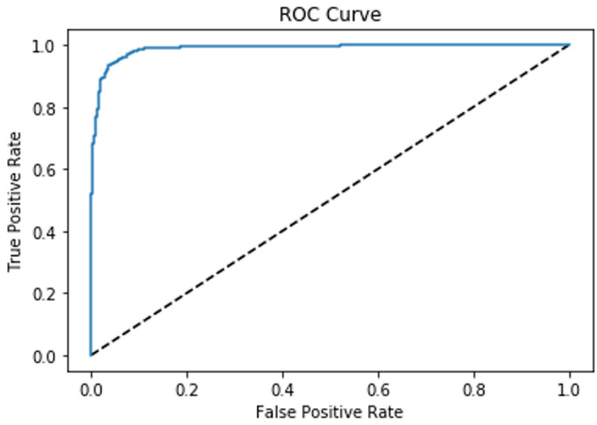


Fig. 6. ROC Curve between True Positive Rate and False Positive Rate.

5 Discussion

The proposed model has a relatively simpler architecture as compared to other architectures that have a relatively deeper network with a large number of training parameters. This is because of the careful design of various hyperparameters and design parameters, which play a vital role in determining the performance of the model. The next section discusses in detail various empirical testing methods used to select

appropriate hypermeters and also a comparison of the performance of our proposed model with various other models developed using similar architecture on the same dataset.

5.1 Empirical Determination of Hyperparameters

The performance of deep learning classifiers is heavily dependent on hyperparameters, and therefore, to achieve the desired result for a particular application, customized models have to be designed.

The first design parameter is the number of CNN layers, in particular, the number of convolutional layers. When the number of convolution layers changes, the performance is significantly impacted because of the distribution of weights and kernels across several layers. The second design parameter is the number of outputs of the FC1 layer. The dense network essentially behaves like a neural network with a hidden layer, and therefore, the number of the output of the FC1 layer plays a vital role in model performance. Therefore, in this paper, we have performed two empirical analyses to select the best performing parameters and results have been plotted as shown in Fig. 7.

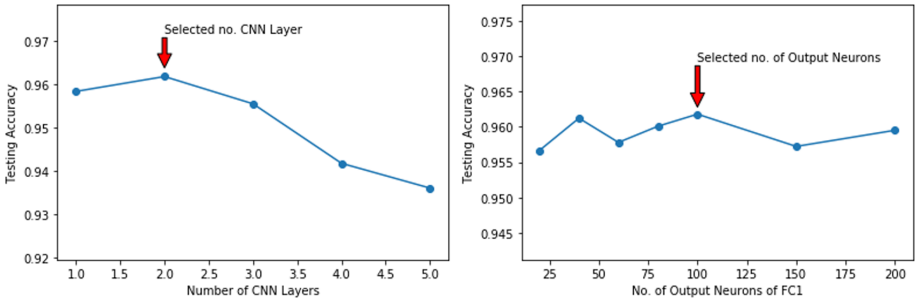


Fig. 7. Plot for empirical testing of design parameters. (a) Testing Accuracy v/s No. of CNN layers and (b) Testing Accuracy v/s No. of Output Neurons of FC1 layer

5.2 Performance Comparison

Next, we compared our classifier model with other models, which also uses the same dataset [15] using the same number of Chest X-ray images as shown in Table 5. The validation accuracy achieved of the proposed model, despite relatively simpler architecture without any pretraining, is 96.18%, which is higher than the accuracy achieved by various models developed using the same dataset.

Table 5. Performance Comparison of the model evaluated on the same dataset

Performance metric	No. of chest X-ray image used	Architecture	Validation accuracy (%)
D. S. Kerman etc. [10]	5856	Pretrained TensorFlow Inception V3 Architecture	92.8
O. Stephen etc. [11]	5856	4 Convolutional and 2 Dense layers	93.73
A. A. Saraiva etc. [12]	5856	7 Convolutional and 3 Dense layers	95.30
Proposed Model	**5856**	**2 Convolutional and 2 Dense layers**	**96.18**

6 Conclusion

In this paper, we proposed the model a nine-layer convolutional neural network for detecting pneumonia from chest X-ray images. The proposed model achieved a training accuracy of 100% (trained for 200 epochs) and a validation accuracy of 96.18% for this problem of binary classification. The result shows that the proposed approach offers a very high prediction accuracy on the chest X-ray images with minimum training parameters. The proposed method can be extended as a generalized technique to assist medical professionals for faster diagnosis of other diseases as well. The future research work will be focusing on to extend this classification model for multiple classes of diseases that are possible to be diagnosed from such similar chest X-ray image datasets.

Acknowledgment. The research work was supported by the Electronics Engineering Department, S. V. National Institute of Technology, Surat, India. We appreciate the effort kept to collect and share the labeled Chest X-ray image dataset [15]. We would also like to appreciate Apurva Randeria's (SVNIT, Surat) effort for developing an interpretable graphical model for the proposed CNN based classifier.

References

1. Ruuskanen, O., Lahti, E., Jennings, L.C., Murdoch, D.R.: Viral pneumonia. Lancet (2011)
2. Islam, S.R., Maity, S.P., Ray, A.K., Mandal, M.: Automatic detection of pneumonia on compressed sensing images using deep learning. In: IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) (2019)
3. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Esteva, A., Kuprel, B., Roberto, A.N., Ko, J., Susan, M.S., Helen, M.B., Thrun, S.: Dermatologist-level classification of skin cancer. Nature (2017)
5. El-Solh, A.A., Hsiao, C.B., Goodnough, S., Serghani, J., Grant, B.J.: Predicting active pulmonary tuberculosis using an artificial neural network (1999)
6. Er, O., Yumusak, N., Temurtas, F.: Chest diseases diagnosis using artificial neural networks. Expert System with Application **37**(12), 7648–7655 (2010)

7. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning, arXiv (2017)
8. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE - Computer Vision and Pattern Recognition (2017)
9. Xia, Y., Wang, H.: ChestNet: a deep neural network for classification of thoracic diseases on chest radiography. ArXiv (2018)
10. Kermany, D.S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **72**(5), 1122–1131 (2018)
11. Stephen, O., Sain, M., Maduh, U.J., Jeong, D.-U.: An efficient deep learning approach to Pneumonia classification in healthcare. J. Healthcare Eng. **2019** (2019). Article ID 4180949
12. Saraiva, A.A., et al.: Classification of images of childhood pneumonia using convolutional neural networks. In: 6th International Conference on Bioimaging (2019)
13. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (2015)
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
15. Kermany, D., Zhang, K., Goldbaum, M.: Labeled Optical Coherence Tomography (OCT) and chest X-Ray images for classification. Mendeley Data (2018)
16. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: NIPS 2018 Proceedings of the 32nd International Conference on Neural Information Processing Systems (2018)
17. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
18. Zhu, W., Zeng, N., Wang, N.: Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. Health Care and Life Science **19**, 67 (2010)
19. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE **10**(3), e0118432 (2015)
20. Brzezinski, D., Stefanowski, J.: Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. Knowledge and Information Systems, 52(2), 531–562 (2017)
21. Junge, M.R.J., Dettori, J.R.: ROC Solid: Receiver Operator Characteristic (ROC) curves as a foundation for better diagnostic tests. Global Spine J. **8**(4), 424–429 (2018)