A convolutional neural network ensemble model for pneumonia detection using chest X-ray images

Sahal Saeed | Fardeen Farhat | Zain Ul Wahab

***Abstract* —** This study provides an improved convolutional neural network (CNN) ensemble model for detecting pneumonia using chest X-ray pictures. The system enhances classification performance and lowers bias toward a single class by merging models with 3x3, 5x5, and 7x7 kernels, followed by the addition of an attention mechanism. The final ensemble, weighted by validation performance, outperformed the original model in terms of recall and F1-scores, indicating its potential as a trustworthy diagnostic aid, particularly in low-resource healthcare settings.

I. Summary

This research paper called "A convolutional Neural network ensemble model for pneumonia detection using chest X-ray images" is about using a special type of computer learning network called a convolutional neural network to help find pneumonia in chest X-ray pictures. The authors created a system that uses a combination of these networks, known as an ensemble model, to look at the X-rays. Multiple models of kernels with sizes 3x3, 5x5 and 7x7 were combined into an ensemble model. By combining the results from these networks, the system can spot pneumonia really well. The goal is to make a tool that can be used easily in places where they might not have a lot of fancy medical equipment.

II. Impementation

My implementation included a lot of testing and training of different models. First of all, I made sure to copy the exact architecture of the model that the original authors used. They used a neural network with 3 convolutional layers. The activation function used on each layer was called Relu. Relu is one of the most useful and most common activation functions in neural networks as it adds non linearity to our model. Specifically, what it does is that it makes negative values equal to zero. On every layer batch normalization was done and max pooling was also done with a pool size of 2x2 to reduce the dimensionality of the data.

At the output layer, sigmoid was used and the threshold values were tweaked each. Below is the threshold values used for all the three models trained.

Table 1.1 Kernal vs Threshold

Kernal	Threshold
3x3	0.875
5x5	0.35
7x7	0.25

That's the model architecture done. The data was prepared before feeding it to the model. I made sure the images were in greyscale. They were also normalized between 0 and 1. Batch size of 32 was used and image sizes were reduced to 180x180. Due to our train set being very unbalanced we decided to use data augmentation added random flips, random rotations, random zoom, random translations and made sure the data was equally distributed.

There were three splits made to feed our model. First of all, the training set which was used to train the model. Secondly, the validation set which was used to fine tune the hyperparameters. Finally, the test set was used to find the evaluation metrics such as accuracy, f1 score, recall and precision.

III. Results achieved

Each model was trained for 10 epochs each. The 3x3 kernel gave us an f1-score on the validation set equal to 0.8889. The f1-score on the test set was equal to 0.9056. The model gave us an accuracy of 0.88.

Table 1.2 Classification Report of 3x3 kernel

	Precision	Recall	F1-score	Support
NORMAL	0.83	0.86	0.85	234
PNEUMONIA	0.91	0.9	0.91	390
accuracy			0.88	624
macro avg	0.87	0.88	0.88	624
weighted avg	0.88	0.88	0.88	624

The 5x5 kernel gave us an f1-score on the validation set equal to 0.6154. The f1-score on the test set was equal to 0.8443. The model gave us an accuracy of 0.82.

Table 2.1 Classification Report of 5x5 kernel

	Precision	Recall	F1-score	Support
NORMAL	0.71	0.86	0.78	234
PNEUMONIA	0.9	0.79	0.84	390
accuracy			0.82	624
macro avg	0.81	0.83	0.81	624
weighted avg	0.83	0.82	0.82	624

The 7x7 kernel gave us an f1-score on the validation set equal to 0.2222. The f1-score on the test set was equal to 0.5760. The model gave us an accuracy of 0.62.

Table 3.1 Classification Report of 7x7 kernel

	Precision	Recall	F1-score	Support
NORMAL	0.5	0.97	0.66	234
PNEUMONIA	0.95	0.41	0.58	390
accuracy			0.62	624
macro avg	0.72	0.69	0.62	624
weighted avg	0.78	0.62	0.61	624

IV. Interpretation of results

The performance metrics reveal distinct characteristics for each kernel size in the pneumonia detection task. The 3x3 kernel model demonstrates a balanced performance with relatively high precision (0.83 for

NORMAL, 0.91 for PNEUMONIA) and recall (0.86 for NORMAL, 0.90 for PNEUMONIA), indicating its ability to accurately classify both normal and pneumonia cases. In contrast, the 5x5 kernel model shows a trade-off, maintaining high precision for PNEUMONIA (0.90) but with a reduced recall for PNEUMONIA (0.79) and precision for NORMAL (0.71), suggesting it's better at identifying pneumonia but less accurate in classifying normal cases. The 7x7 kernel model exhibits a more pronounced imbalance, with very high recall for NORMAL (0.97) but low recall for PNEUMONIA (0.41) and precision for NORMAL (0.50), indicating a strong tendency to classify cases as normal, even when they are not. Overall, the 5x5 kernel model was weighted the most for the ensemble model because it had the highest recall.

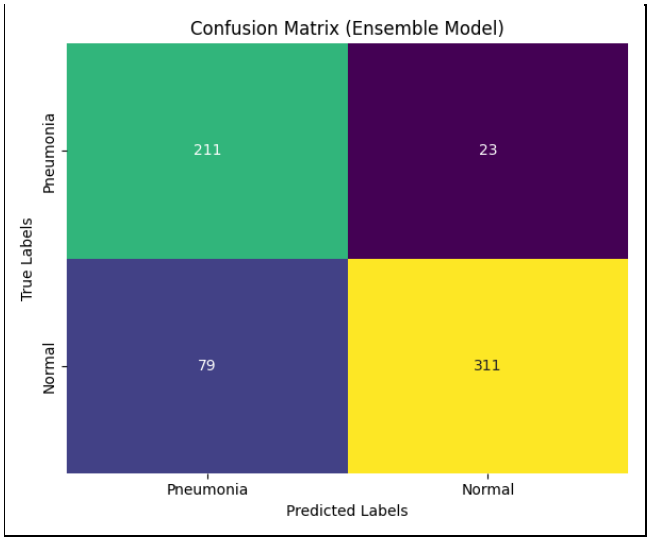
V. Ensemble Model Results

The ensemble model was then evaluated. First of all, we need to check how we found the ensemble model's calculation. Weights were given to each model's probability. 3x3 kernel was given the weight 0.3, 5x5 was given the weight 0.6 and 7x7 due to its huge imbalance was given the weight 0.1. Below formula is used to find the weighted sum.

$$Weighted\ Sum = \Sigma\ (weight_i * prediction_i)$$

Below are the ensembel model's results

Metric	Value
Accuracy	0.8365
Precision	0.9311
Recall	0.7974
F1_score	0.8591



VI. Contribution

Due to our model being very biased towards a single class even though we have used data augmentation, in our phase 2 we decided to go for a model enhancement approach. We decided to add an attention mechanism to our convolutional neural network. We decided to keep the other architecture i.e. the number of layers the same but we changed one thing. An attention mechanism is applied after the third convolutional block. A 1x1 convolution with sigmoid activation generates an attention map. This map is element-wise multiplied with the feature map to enhance important regions.

This model enhancement would in theory help us bridge this gap and make sure that our model does not result in false negative and false positive predictions. I have also added the mechanism that it decides an optimal threshold for our sigmoid for each kernel size on the basis of our validation set. Let us move onto the results.

VII. Results After Contribution

For the 3x3 kernel we got an amazing f1-score on the validation set of 0.9412 and on the test set we got 0.8723. We got an accuracy of 0.83. This shows us that our new technique is working.

	Precision	Recall	F1-score	Support
NORMAL	0.88	0.63	0.73	234
PNEUMONIA	0.81	0.95	0.87	390
accuracy			0.83	624
macro avg	0.84	0.79	0.8	624
weighted avg	0.83	0.83	0.82	624

For the 5x5 kernel we got an f1-score on the validation set of 0.6957 and on the test set we got 0.7808. We got an accuracy of 0.65.

	Precision	Recall	F1-score	Support
NORMAL	1	0.06	0.12	234
PNEUMONIA	0.64	1	0.78	390
accuracy			0.65	624
macro avg	0.82	0.53	0.45	624
weighted avg	0.78	0.65	0.53	624

For the 7x7 kernel we got an f1-score on the validation set of 0.0 and on the test set we got 0.7241. We got an accuracy of 0.67.

	Precision	Recall	F1-score	Support
NORMAL	0.55	0.61	0.58	234
PNEUMONIA	0.75	0.7	0.72	390
accuracy			0.67	624
macro avg	0.65	0.66	0.65	624
weighted avg	0.68	0.67	0.67	624

VIII. Interpretation of Results After Contribution

As we can see in the above results our 7x7 kernel model is performing terribly. There is no need to use our 7x7 kernel due to its very bad f1-score i.e. 0.0 on our validation set. We decided to go for the below ensemble weights for each kernel size.

Model	Weight
3x3	0.575
5x5	0.425
7x7	0

IX. Final Ensemble Model Results

Using the weights discussed in the previous paragraph, we decided to take the weighted sum using the formula:

$$Weighted\ Sum = \Sigma\ (weight_i * prediction_i)$$

The results we got were very satisfactory and helped us reach our best results values, beating the original author's results.

Table 4 Final ensemble model results

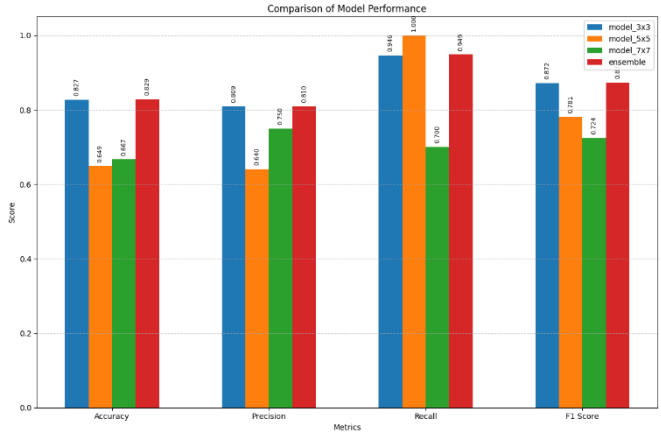
Metric	Value
F1 Score (Validation)	0.9412
F1 Score (Test)	0.8737
Accuracy (Test)	0.8285
Precision (Test)	0.8096
Recall (Test)	0.9487

X. Comparison with the original results

Overall, our model has gotten much better in learning intricate patterns. Its overall biasness has reduced by a great margin as well. Below is the side-by-side comparison:

Metric	Old Value	New Value	Change	Percentage Change
Accuracy (Test)	0.8365	0.8285	-0.008	-0.96%
Recall (Test)	0.7974	0.9487	0.1513	18.97%
F1 Score (Test)	0.8591	0.8737	0.0146	1.70%

Below is also the comparison of the models of the kernels and the ensemble model with our new architectural change.



XI. Conclusion

In conclusion, this new ensemble model of ours is a big step ahead from the original model. It can now handle the ability to learn intricate patterns. It can also now learn to not be biased towards a specific class due to its attention mechanism. One of the biggest advantages of this new attention mechanism is the fact that it can learn and analyze intricate details in an image which are very important for X-ray images.

XII. Future work

In the future, transfer learning could be taken in consideration, as for now we have not used any pre trained convolutional neural network like Resnet50 or Imagenet. Transfer learning could benefit this methodology a lot.

Another future work could be to add the concept of Convolutional Block Attention Module (CBAM), it learns intricate details much more better then our current implementation.

References

[1] Bhatt, H. and Shah, M., 2023. A Convolutional Neural Network ensemble model for Pneumonia Detection using chest X-ray images. *Healthcare Analytics*, *3*, p.100176.

[2] Soydaner, Derya. "Attention mechanism in neural networks: where it comes and where it goes." Neural Computing and Applications 34.16 (2022): 13371-13385.

[3] Cui, Yiming, et al. "Attention-over-attention neural networks for reading comprehension." arXiv preprint arXiv:1607.04423 (2016).

[4] Chen, Wei, and Ke Shi. "Multi-scale attention convolutional neural network for time series classification." *Neural Networks* 136 (2021): 126-140.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.