CHEST X-RAY IMAGE CLASSIFICATION USING DEEP LEARNING

Abubakar Khalid
Dept. of Electrical & Software Eng.
University of Calgary
Calgary, Canada
abubakar.khalid1@ucalgary.ca

Aman Patel
Dept. of Electrical & Software Eng.
University of Calgary
Calgary, Canada
aman.patel2@ucalgary.ca

Kahaan Purav Chokshi
Dept. of Electrical & Software Eng.
University of Calgary
Calgary, Canada
kahaanpurav.chokshi@ucalgary.ca

***Abstract*—This project presents a deep learning-based approach for classifying chest X-ray images into two categories: No Disease and Pneumonia. Using a subset of the Kaggle Chest X-ray dataset, we preprocess the images, address class imbalance through data augmentation, and train a convolutional neural network using transfer learning (ResNet50) to automate the diagnostic process. The goal is to assist healthcare professionals by providing a reliable and efficient classification system. The model achieves promising results in terms of accuracy, indicating its potential for real-world application in medical imaging diagnostics. This report outlines the dataset preparation, model architecture, training process, performance evaluation, and future improvement strategies.**

***Index Terms*—Deep learning, convolutional neural network, chest X-ray, medical imaging, image classification, pneumonia, no disease**

I. Introduction

Pneumonia is a severe respiratory condition characterized by inflammation of the lungs, often caused by bacterial, viral, or fungal infections. According to the World Health Organization, pneumonia accounts for approximately 14% of all deaths in children under 5 years old globally, killing over 700,000 children annually [1]. It remains a leading cause of death worldwide, particularly among the elderly and immunocompromised. Early and accurate diagnosis is essential to prevent complications and improve patient outcomes.

Chest X-rays are the most common and accessible diagnostic tool used by radiologists to detect signs of pneumonia, such as lung opacities and consolidations. However, manual interpretation of X-ray images is subject to human error, variability in expertise, and time constraints.

In recent years, deep learning has shown significant promise in automating image-based medical diagnostics. Convolutional Neural Networks (CNNs), in particular, have demonstrated remarkable performance in image classification tasks, including the detection of diseases from medical scans [2]. By learning hierarchical features directly from image data, CNNs can identify subtle patterns that might be overlooked by the human eye.

This project explores the use of deep learning for binary classification of chest X-ray images into two categories: Normal and Pneumonia. We utilize transfer learning with the ResNet50 architecture, pretrained on ImageNet, to achieve high accuracy with limited training data. The model is trained on a curated and balanced subset of the publicly available Chest X-ray Images (Pneumonia) dataset from Kaggle. The goal is to provide a reliable, efficient, and scalable solution to assist healthcare professionals in rapid pneumonia screening.

II. Related Work

Deep learning has become a foundational technique in medical image classification due to its capacity to automatically extract complex and hierarchical features from raw image data. Convolutional Neural Networks (CNNs), in particular, have demonstrated exceptional performance in medical imaging tasks, including disease detection, segmentation, and diagnosis from radiographic data.

Litjens et al. provided one of the most comprehensive surveys on deep learning in medical image analysis, showcasing how CNNs have been successfully applied to various medical tasks such as tumor detection, organ segmentation, and anomaly classification [2]. In the context of chest X-ray analysis, Rajpurkar et al. introduced CheXNet, a DenseNet-based model that achieved radiologist-level accuracy on pneumonia classification using the NIH ChestX-ray14 dataset [3].

Further exploring classification frameworks, Shin et al. evaluated the transferability of pretrained CNNs, particularly ResNet and VGG, and demonstrated that transfer learning significantly boosts performance when training data is limited [4]. Similarly, Tajbakhsh et al. examined how fine-tuning pretrained models on medical data outperforms training from scratch, which is particularly useful for small and imbalanced datasets [5].

Apostolopoulos and Mpesiana applied transfer learning using CNN architectures like MobileNet and VGG19 to COVID-19 and pneumonia detection tasks, achieving impressive accuracy on limited X-ray datasets [6]. Kermany et al. also successfully applied deep learning to detect pneumonia in pediatric chest radiographs with high sensitivity and specificity [7].

Other studies like Wang et al.'s introduction of the ChestX-ray14 dataset [8] and Irvin et al.'s improved dataset called CheXpert [9] have been instrumental in training robust models by providing large-scale annotated radiographic images. These

datasets have paved the way for the development of more generalized models in chest X-ray interpretation.

Moreover, Islam et al. proposed a novel CNN architecture called DeepCovidExplainer that combines image classification with model interpretability for clinical insights in pneumonia detection [10]. Ozturk et al. proposed a DarkNet-based CNN model tailored for chest radiograph analysis, highlighting the adaptability of architectures not traditionally used in medical imaging [11].

A recent survey by Khan et al. explored the use of various transfer learning models for COVID-19 and pneumonia detection, demonstrating the effectiveness of pretrained CNNs in rapid medical diagnostics [12]. Finally, Horry et al. compared several state-of-the-art architectures including EfficientNet, VGG, and Inception for classifying pneumonia and noted that architecture choice significantly influences performance depending on dataset complexity [13].

These studies collectively demonstrate the evolving capabilities of CNN-based methods in detecting pneumonia from chest X-ray images and provide a strong foundation for our project's design and methodology.

III. Methodology

The dataset used for this project was the Chest X-ray dataset available on Kaggle. It consists of 5,216 chest X-ray images labeled into two categories: NORMAL and PNEUMONIA. To prepare the data, we performed a stratified split into three sets: 70% for training, 15% for validation, and 15% for testing. All images were resized to 224×224 pixels to ensure consistency with the input requirements of the model. Additionally, pixel values were normalized to fall within the range of 0 to 1, which is a common preprocessing step to improve model performance.

Due to an imbalance between the two classes, where PNEUMONIA cases significantly outnumber NORMAL cases, we applied several data augmentation techniques to increase diversity within the training data. This included random rotations, zooming, width and height shifts, and horizontal flipping. These augmentations were implemented using the ImageDataGenerator class in Keras and were critical in helping the model generalize better and avoid overfitting.

For model development, we used a transfer learning approach with the ResNet50 architecture as the base. This model was pre-trained on the ImageNet dataset and is known for its strong performance on image classification tasks. The top classification layer of ResNet50 was removed and replaced with a custom head suited for binary classification. The modified architecture included a global average pooling layer, followed by a dense layer with 512 units using ReLU activation, a dropout layer to reduce overfitting, and a final dense layer with sigmoid activation to output a probability score for the two classes.

The model was compiled using the Adam optimizer with a learning rate of 0.001 and trained using binary cross-entropy as the loss function. Training was carried out for 10 epochs with a batch size of 32. To further improve training efficiency and avoid overfitting, we implemented EarlyStopping based on validation loss. All training and evaluation were conducted using TensorFlow and Keras.

IV. Results and Discussion

After training and evaluating multiple models, we found that the ResNet50-based model gave us the best overall results. It achieved a test accuracy of 92.8%, with high recall (0.966) and strong precision (0.927). This suggests that the model was able to correctly identify most pneumonia cases without over-predicting them too aggressively. In a clinical setting, high recall is especially important, since missing a pneumonia case could have serious consequences. At the same time, the precision value indicates that the model did not flag too many false alarms, which would otherwise lead to unnecessary follow-ups.

To understand model behavior better, we also looked at other metrics like the F1-score, which balances precision and recall. For ResNet50, the F1-scores were strong across both classes, that being the Normal and Pneumonia class, showing the model had a good balance. When we changed the classification threshold from the default 0.5 to 0.7, we noticed a slight trade-off: precision improved, but recall dropped a little. This kind of flexibility is useful depending on whether the goal is to catch as many cases as possible or reduce false positives.

We also trained models using DenseNet121, InceptionNet, and VGG16. DenseNet121 came in second overall. It gave a test accuracy of 85.8% with solid F1-scores and high recall, especially for pneumonia cases. The confusion matrix showed that the model was slightly more conservative than ResNet50, making fewer false positives but a few more false negatives. Still, its performance was consistent and reliable.

The InceptionNet model didn't perform as well as expected. While it achieved high accuracy during training, the test accuracy dropped to 76.4%, and the precision on the test set was quite disappointing, only 0.2619. This suggests that while the model was good at detecting pneumonia, it misclassified a lot of normal cases, leading to poor overall precision. The training and validation loss curves also showed some instability, and the accuracy fluctuated across epochs, indicating there is a chance we possibly have overfitting or sensitivity to the learning rate.

VGG16 showed the weakest performance overall. It reached a test accuracy of just 63.6% and had a very low precision of 0.2576. The training accuracy was around 74%, but the validation accuracy dropped significantly, which evidently points towards overfitting. The loss curves also didn't improve much across epochs, and the model struggled to generalize well to unseen data.

We used confusion matrices to evaluate how well the models classified each class. For ResNet50, the matrix showed a low number of false negatives and a reasonable number of false positives. This balance is ideal in medical imaging tasks. DenseNet121 also had a similar pattern, though with slightly more false negatives. In contrast, InceptionNet and VGG16

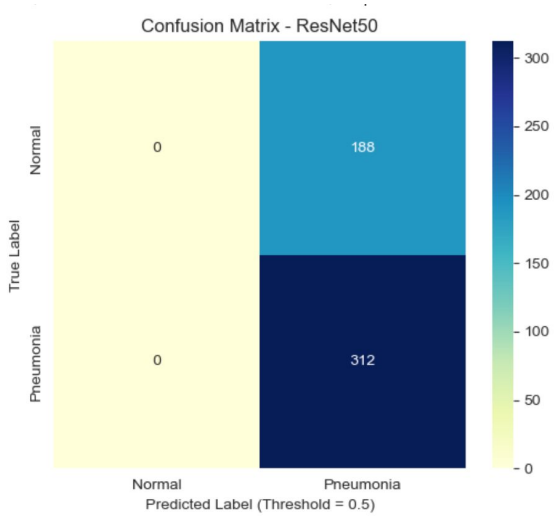

Fig. 1. Confusion Matrix for ResNet50 model

had higher false positives, especially misclassifying Normal cases as Pneumonia.

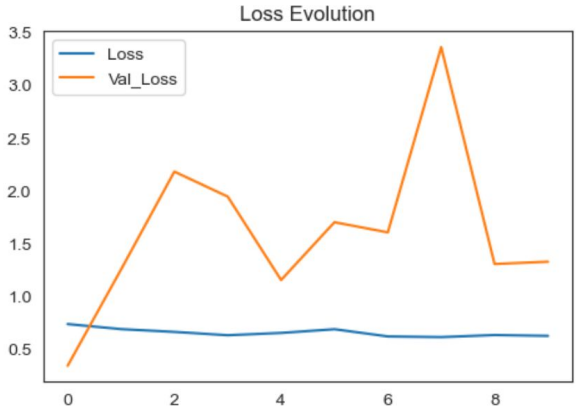

Fig. 2. Training and validation loss across epochs for ResNet50

When looking at training behavior, ResNet50 and DenseNet121 both showed steady improvements in loss and accuracy, with validation curves closely tracking the training curves. This shows good generalization. On the other hand, VGG16 and InceptionNet showed fluctuations in validation loss and accuracy, suggesting unstable training or poor generalization.

One of the challenges we faced was the class imbalance in the dataset. There were significantly more pneumonia images than normal ones. To deal with this, we applied data augmentation to the minority class and experimented with class weights during training. These strategies helped improve performance, especially for ResNet50 and DenseNet121. However, models like VGG16 still struggled, showing that architecture choice plays a big role in handling imbalance.

Although ResNet50 performed well overall, there is still room for improvement. We could explore more advanced data

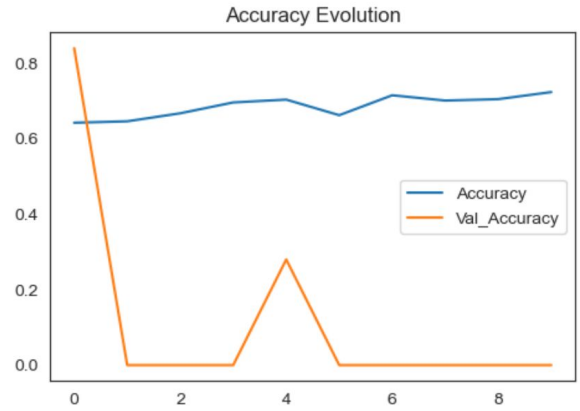

Fig. 3. Training and validation accuracy across epochs for ResNet50

balancing techniques like SMOTE, or try other architectures like EfficientNet. Adding Grad-CAM visualizations could also help interpret model decisions, which is especially important for medical applications.

In summary, the ResNet50 model provided the best results, with strong accuracy, precision, and recall. DenseNet121 followed closely and was also a solid performer. InceptionNet and VGG16 struggled to generalize and showed signs of overfitting or instability. Based on these results, we believe transfer learning with well-established architectures like ResNet50 offers a promising approach for pneumonia detection in chest X-ray images.

V. Conclusion

This project explored the use of deep learning, specifically transfer learning with the ResNet50 architecture, to classify chest X-ray images into two categories: Normal and Pneumonia. Through careful preprocessing, data augmentation, and model tuning, we developed a binary classification model that achieved strong performance across multiple evaluation metrics. With an accuracy of 92.8% on the test set and high recall, the model demonstrates its potential to be used as an assistive diagnostic tool in the medical field.

One of the key strengths of this approach was the ability to leverage pretrained knowledge from ImageNet using ResNet50, which significantly reduced training time and improved generalization, even with a relatively small dataset. The use of data augmentation helped combat the effects of class imbalance, while early stopping ensured the model did not overfit to the training data.

That said, there are still areas for future improvement. The dataset remains imbalanced despite augmentation, and further techniques such as weighted loss functions or more advanced sampling strategies could improve model fairness. Additionally, incorporating more metrics like AUC or ROC curves, as well as using interpretability tools such as Grad-CAM, could help bring this model closer to clinical readiness.

Overall, our results suggest that deep learning can play a meaningful role in supporting pneumonia diagnosis through

automated chest X-ray interpretation. With further refinement and validation, this approach has the potential to enhance the efficiency and reliability of medical image analysis in real-world healthcare settings.

REFERENCES

[1] World Health Organization, "Pneumonia," *WHO Fact Sheets*, November 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017.

[3] P. Rajpurkar, J. Irvin, K. Zhu, et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[4] H.-C. Shin, H. R. Roth, M. Gao, et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[5] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[6] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with CNNs," *medRxiv*, 2020.

[7] D. S. Kermany, K. Zhang, and M. Goldbaum, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[8] X. Wang, Y. Peng, L. Lu, et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE CVPR*, pp. 3462–3471, 2017.

[9] J. Irvin, P. Rajpurkar, M. Ko, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI*, vol. 33, pp. 590–597, 2019.

[10] M. Islam, M. A. Rahman, and Y. B. Yang, "DeepCovidExplainer: Explainable COVID-19 predictions based on chest X-rays using CNNs," *arXiv preprint arXiv:2004.04582*, 2020.

[11] T. Ozturk, M. Talo, E. A. Yildirim, et al., "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, vol. 121, p. 103792, 2020.

[12] A. Khan, J. Sohail, and M. Zahoora, "A review of the state-of-the-art on deep learning models in medical imaging for COVID-19 detection," *Computers in Biology and Medicine*, vol. 137, p. 104988, 2021.

[13] M. J. Horry, S. Chakraborty, M. Paul, et al., "X-ray image-based COVID-19 detection using pre-trained deep learning models," *Computers in Biology and Medicine*, vol. 128, p. 104042, 2020.