

Comparative Analysis of VGG16 and InceptionV3 for Multi-Label Chest X-Ray Classification

João Victor Morgado, Virgílio de Melo Langoni, Carlos Rafael Nogueira da Silva

Abstract—This paper employs the NIH ChestX-ray14 dataset to compare the VGG16 and InceptionV3 architectures for multi-label chest X-ray classification. The impact of CLAHE-based contrast enhancement on convolutional neural network performance is investigated. Transfer learning was applied using ImageNet-pretrained weights. ROC curves and AUC metrics were used to assess classification performance, highlighting the influence of preprocessing and architectural design on medical image analysis. The results show that InceptionV3 outperforms VGG16, while CLAHE-based preprocessing improves the classification performance of both architectures.

Keywords—Deep Learning, Chest X-ray classification, VGG16, InceptionV3, CLAHE.

I. INTRODUCTION

Deep learning techniques have significantly advanced computer vision applications in medical imaging [1], [2], particularly for chest radiography-based Computer-Aided Diagnosis (CAD) systems. Convolutional Neural Networks (CNNs) have demonstrated remarkable capability to automatically extract complex visual features directly from raw pixel data, thus minimizing the need for hand-crafted feature extraction. Among publicly available medical imaging repositories, the NIH ChestX-ray14 dataset [3] has become a benchmark for evaluating deep learning models due to its large scale and multi-label nature. However, this dataset presents significant challenges, including severe class imbalance and substantial overlap among disease patterns, which can hinder model generalization.

Recent studies have shown that image preprocessing techniques can significantly enhance CNN performance in radiographic analysis. In particular, Contrast Limited Adaptive Histogram Equalization (CLAHE) [4] has been widely employed to improve local contrast and sharpen edge definition, enabling the extraction of subtle pathological features.

In this context, this study conducts a comparative analysis between VGG16 [5] and InceptionV3 [6], both fine-tuned via transfer learning, for multi-label chest X-ray classification. Specifically, the impact of CLAHE preprocessing on classification performance is evaluated through Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metrics.

João Victor Morgado, Virgílio de Melo Langoni and Carlos Rafael Nogueira da Silva are with the Federal University of Triângulo Mineiro (UFTM), Uberaba-MG, Brazil. E-mails: d202310294@uftm.edu.br, virgilio.langoni@uftm.edu.br, carlos.nogueira@uftm.edu.br. This work was supported by CNPq.

II. METHODOLOGY

This section presents the methodological framework adopted in this study, including the dataset description, image preprocessing procedure, preprocessing evaluation metrics, CNN architectures, training strategy, and classification performance metrics used throughout the experimental analysis.

A. Dataset

The experiments were conducted using the NIH ChestX-ray14 dataset [3], which contains more than 100,000 chest radiographs associated with 14 thoracic pathologies. Due to the multi-label nature of the dataset, a single image may simultaneously present multiple pulmonary abnormalities.

In this work, only PA-view (posteroanterior) radiographs were considered to ensure geometric consistency during training. Additionally, the majority “No Finding” class was randomly under-sampled to reduce class imbalance and mitigate classification bias. To reduce multi-label complexity, only images containing up to two simultaneous pathologies were selected.

To further address class imbalance, minority classes such as Hernia and Edema were balanced through random oversampling during the training stage by replicating underrepresented samples. This strategy increased the representation of rare pathologies during optimization, reducing bias toward majority classes.

B. Image Preprocessing

All chest radiographs were resized to the native input dimensions of each architecture (224×224 for VGG16 and 299×299 for InceptionV3). Furthermore, pixel intensities were normalized according to the preprocessing requirements of each pre-trained network.

To enhance local contrast and improve the visibility of subtle anatomical structures in chest radiographs, CLAHE was employed. As illustrated in Fig. 1, the preprocessing procedure improved edge definition without excessive noise amplification. Following preprocessing, both entropy and Laplacian variance increased considerably. These metrics, which correlate with image sharpness and information content, indicate that local contrast and edge visibility were enhanced.

C. Preprocessing Evaluation Metrics

Two image quality metrics were used to quantify the influence of CLAHE preprocessing: Shannon entropy and Laplacian variance [7]. Shannon entropy was used to quantify information content and grayscale diversity, while Laplacian

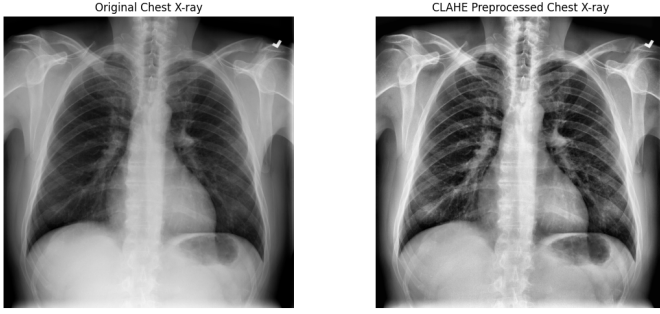


Fig. 1. Comparison between original and CLAHE-preprocessed chest radiographs.

variance was used to estimate image sharpness and edge enhancement.

Shannon entropy is defined as

$$H = - \sum_{i=0}^{L-1} p(i) \log_2 p(i), \quad (1)$$

where $p(i)$ represents the probability distribution of grayscale intensity level i .

The variance of the Laplacian operator is defined as

$$\sigma_L^2 = \frac{1}{N} \sum_{i=1}^N (L_i - \mu_L)^2, \quad (2)$$

where L_i denotes the Laplacian response at pixel i , μ_L is the mean Laplacian value, and N is the total number of pixels.

These measurements were calculated before and after pre-processing to evaluate the impact of CLAHE on image sharpness, contrast enhancement, and information content.

D. CNN Architectures

Deep learning architectures have been extensively applied to chest X-ray analysis in recent years. CNN-based approaches such as VGGNet, ResNet, DenseNet, and Inception architectures have demonstrated strong performance in pulmonary disease detection tasks [5], [6]. In this work, two convolutional neural network architectures were evaluated: VGG16 and InceptionV3. Both models were trained using transfer learning from ImageNet pre-trained weights.

The VGG16 architecture is recognized for its sequential convolutional structure and ability to extract fine texture patterns from medical images [5]. Fig. 2 illustrates a simplified representation of the VGG16 architecture, where convolutional layers are sequentially stacked to progressively extract low- and high-level visual features.

On the other hand, InceptionV3 employs multi-scale convolutional modules capable of extracting features at different spatial resolutions while maintaining reduced computational complexity [6]. As illustrated in Fig. 3, the architecture processes feature maps through parallel convolutional paths with different kernel sizes.

To support multi-label classification of 14 thoracic diseases, custom fully connected layers composed of dense, dropout, and sigmoid activation layers were added to the original architectures.

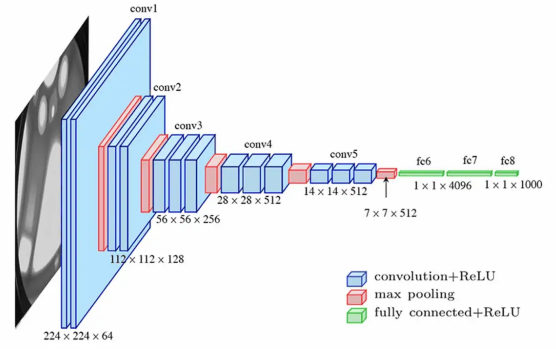


Fig. 2. Simplified representation of the VGG16 architecture. Source: [8].

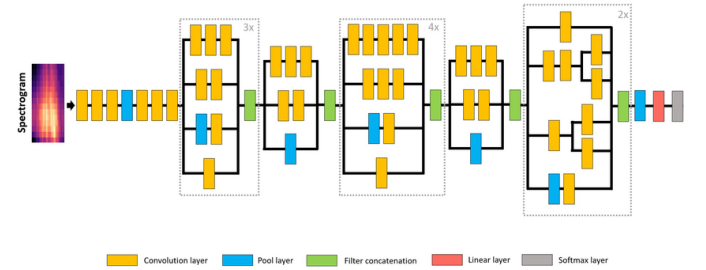


Fig. 3. Simplified representation of the InceptionV3 architecture. Source: [9].

E. Training Strategy

During the first training stage, only the classification head was trained while the convolutional backbone remained frozen. Subsequently, fine-tuning was performed by unfreezing the deeper convolutional blocks of each architecture using a lower learning rate to adapt the feature extraction process to chest radiograph patterns.

Binary cross-entropy is widely used in multi-label classification tasks due to its probabilistic interpretation and optimization stability [10]. Due to the severe class imbalance present in the NIH ChestX-ray14 dataset, a weighted binary cross-entropy loss function was adopted during training to mitigate bias toward majority classes.

The weighted loss function is defined as

$$\mathcal{L} = - \frac{1}{N} \sum_{i=1}^N (w_i y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)), \quad (3)$$

where y_i represents the ground-truth label, $\hat{y}_i$ is the predicted probability, and w_i denotes the positive class weight associated with each pathology class.

F. Classification Performance Metrics

To evaluate the classification performance of the proposed models, ROC curves and the AUC metric were employed. ROC analysis enables the assessment of the trade-off between sensitivity and specificity across different decision thresholds, while AUC provides a quantitative measure of the overall discriminative capability of the classifier. These metrics are particularly suitable for multi-label and imbalanced medical imaging datasets, where classification performance must be evaluated independently of a fixed threshold.

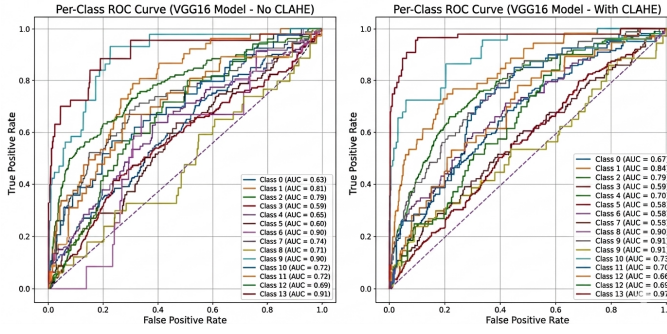


Fig. 4. Comparison of per-class ROC curves obtained using the VGG16 architecture before and after CLAHE preprocessing.

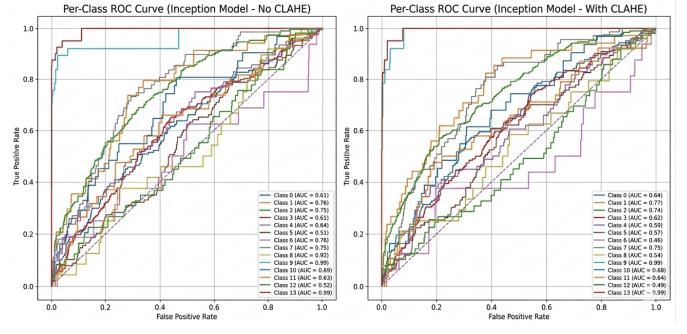


Fig. 5. Comparison of per-class ROC curves obtained using the InceptionV3 architecture before and after CLAHE preprocessing.

III. EXPERIMENTAL RESULTS

A. Impact of CLAHE Enhancement

The CLAHE preprocessing stage produced considerable improvements in both image sharpness and local contrast. Quantitative analysis revealed substantial increases in Shannon entropy and Laplacian variance after preprocessing, indicating enhanced edge definition and improved representation of local image details in chest radiographs.

Table I summarizes the average entropy and Laplacian variance values obtained before and after preprocessing.

TABLE I

AVERAGE ENTROPY AND LAPLACIAN VARIANCE BEFORE AND AFTER CLAHE PREPROCESSING.

Metric	Original	CLAHE	Gain (%)
Entropy	7.4613	7.6215	+2.15%
Laplacian Variance	373.22	938.29	+151.40%

The increase in entropy suggests a more uniform grayscale intensity distribution and greater information richness after preprocessing. Additionally, the considerable increase in Laplacian variance indicates substantial enhancement of high-frequency anatomical structures and edge definition.

These improvements are particularly relevant for chest X-ray analysis, since subtle pulmonary abnormalities often exhibit low local contrast and poorly defined boundaries.

B. ROC Curve Analysis

Figures 4 and 5 present the ROC curves obtained for the evaluated CNN architectures under both original and CLAHE-preprocessed conditions. CLAHE preprocessing enhanced local contrast, emphasizing relevant radiographic features and improving the classification performance of both models. The ability to extract multi-scale features allowed InceptionV3 to achieve superior classification performance compared to VGG16.

The multi-scale feature extraction capability is particularly important in chest radiography analysis, since pathological findings can exhibit significant variations in size, shape, and intensity distribution.

IV. CONCLUSION

This work presented a comparative analysis between the VGG16 and InceptionV3 architectures for multi-label chest

X-ray classification using the NIH ChestX-ray14 dataset. The experimental results demonstrated that CLAHE preprocessing considerably improved image quality, as evidenced by increased entropy and Laplacian variance, indicating enhanced local contrast and sharper anatomical structures, which in turn positively impacted classification performance. Both CNN models benefited from this preprocessing stage, with InceptionV3 achieving superior performance, particularly in low-contrast pathological patterns, due to its multi-scale feature extraction capability.

Overall, the obtained results suggest that combining CLAHE preprocessing with transfer learning-based CNN architectures represents a promising approach for computer-aided diagnosis systems in chest radiography. Future work will involve evaluating additional deep learning architectures, data augmentation techniques, and attention mechanisms for improved pathological feature localization.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.
- [4] K. Zuiderveld, *Contrast limited adaptive histogram equalization*. San Diego, CA, USA: Academic Press Professional, Inc., 1994, p. 474–485.
- [5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, May 2015.
- [6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Dec. 2016, pp. 2818–2826.
- [7] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 4th ed. New York, NY: Pearson, 2018.
- [8] M. Ferguson, R. Ak, Y.-T. T. Lee, and K. H. Law, "Automatic localization of casting defects with convolutional neural networks," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 1726–1735.
- [9] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Transformers for Urban Sound Classification—A Comprehensive Performance Evaluation," *Sensors*, vol. 22, no. 22, 2022.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.