Cardiomegaly Detection in Chest X-rays using a Convolutional Neural Network: Implementation and Comparative Analysis

Vatsal Roy
dept. Computer Science (Of Long Island University, Brooklyn)

Kunj Patel
dept. Computer Science (Of Long Island University, Brooklyn)

Riya Gardharia
dept. Computer Science (Of Long Island University, Brooklyn)

Shreya Vaghela
dept. Computer Science (Of Long Island University, Brooklyn)

***Abstract*—This paper presents the development and evaluation of a custom convolutional neural network (CNN) for the automated detection of cardiomegaly from chest X-ray images. Cardiomegaly, or heart enlargement, is a critical indicator of cardiovascular disease, a leading cause of global mortality. Addressing the need for precise and scalable diagnostic tools, we designed and trained a custom CNN model on a curated dataset of 5,447 chest X-rays sourced from the NIH ChexPert collection. The model underwent rigorous training with regularization techniques, including early stopping and adaptive learning rate reduction, to optimize performance and mitigate overfitting. On an unseen test set, the model achieved an overall accuracy of 79.8% and demonstrated strong discriminative capability with an Area Under the Curve (AUC) of 0.906. However, a detailed analysis of its performance revealed a significant limitation in its sensitivity, with a recall of only 71% for identifying positive cardiomegaly cases. When benchmarked against state-of-the-art architectures from the literature, which report accuracies in the 94–99.8% range, our custom model serves as a solid baseline but underscores the performance gap that must be addressed for clinical viability.**

***Index Terms*—Cardiomegaly, Chest X-ray, Computer-aided diagnosis, Convolutional neural networks, Deep learning, Medical imaging.**

I. Introduction

Cardiomegaly, the abnormal enlargement of the heart, is a crucial radiological marker for a spectrum of cardiovascular diseases (CVDs), which are collectively responsible for approximately one-third of all global deaths annually. Early and accurate detection of this condition is therefore vital for timely clinical intervention and improved patient outcomes. The chest X-ray (CXR) remains a cornerstone of thoracic imaging due to its cost-effectiveness and accessibility. Traditionally, clinicians diagnose cardiomegaly by visually assessing the CXR and calculating the Cardiothoracic Ratio (CTR), where a value exceeding 0.50 typically indicates enlargement. However, this manual method is fraught with limitations, including significant inter-observer variability dependent on clinical experience, the time-consuming nature of the task, and potential for diagnostic delays in high-volume settings.

The proliferation of Artificial Intelligence (AI), particularly deep learning algorithms like Convolutional Neural Networks (CNNs), presents a transformative opportunity to overcome these challenges. CNNs can analyze medical images with high precision, identifying complex spatial patterns associated with pathologies that may be subtle to the human eye. The motivation for this project is to harness this potential to create an automated, precise, and scalable diagnostic tool for cardiomegaly screening. Such a system could serve as an efficient decision-support tool, reducing variability and expediting the diagnostic process.

The primary objective of this paper is to detail the development, training, and rigorous evaluation of a custom-designed CNN model for the binary classification of cardiomegaly from CXRs. Furthermore, we benchmark the performance of our model against established findings from contemporary research to contextualize its efficacy and identify areas for improvement.

II. Literature Review

A strategic review of existing literature is crucial for contextualizing the performance of any new model. This section summarizes the methodologies and benchmark results of three pertinent studies in automated cardiomegaly detection.

Sarpotdar [1] developed a deep learning model for cardiomegaly detection using a customized U-Net architecture. The model was trained on the ChestX-ray8 dataset, an open-source collection of chest radiographs. The study reported impressive performance metrics, achieving a diagnostic accuracy of 94%, a sensitivity of 96.2%, and a specificity of 92.5%, demonstrating the power of specialized segmentation-based networks in this domain.

Zhu et al. [2] designed a comprehensive computer-assisted diagnosis (CAD) pipeline that employs a segmentation-based approach to automatically calculate the Cardiothoracic Ratio (referred to as CV-CTR) for classification. Their system, evaluated on the large-scale MIMIC-CXR dataset, achieved a binary classification accuracy of 95.2% with a corresponding Area Under the Curve (AUC) of 0.95. This work highlights the effectiveness of integrating traditional clinical metrics into an automated deep learning workflow.

In a comparative study, Ayalew et al. [3] evaluated several transfer learning models for early-stage cardiomegaly detection. Using a dataset of 4400 images, their research identified ResNet-50 as the top-performing architecture. The ResNet-50 model delivered exceptional results, achieving a test accuracy of 99.8% with perfect precision (100%) and recall (100%). This study underscores the significant advantage of leveraging powerful, pre-trained architectures for medical imaging tasks.

Collectively, these studies indicate that state-of-the-art approaches utilizing advanced architectures like U-Net and ResNet-50 consistently achieve accuracies well above 90%. This sets a high performance benchmark for the custom CNN model developed in this project.

III. Methodology

This section provides a granular, step-by-step account of the machine learning pipeline implemented in this study.

A. Data Collection and Preparation

The foundation of this study is a specialized dataset derived from the NIH ChexPert collection, which contains 112,120 chest X-ray images. We filtered this extensive database to create a focused subset of 5,544 images containing only positive or negative labels for cardiomegaly. During an initial quality assurance check, we discovered that 97 of these images contained missing pixel values (NaNs) and were subsequently removed. This critical step resulted in a final, clean dataset of 5,447 images for our experiments. To ensure efficient data handling and reduce computational overhead during training, the preprocessed images and their corresponding binary labels were serialized and stored in a single HDF5 file.

B. Image Preprocessing and Normalization

To prepare the images for input into the neural network, a two-step preprocessing pipeline was applied to each radiograph:

- **Resizing:** All images were standardized to a uniform resolution of 224×224 pixels using OpenCV's bilinear interpolation functionality.
- **Normalization:** Pixel intensity values were scaled from their original 8-bit integer range of [0, 255] to a floating-point range of [0, 1] by dividing each pixel value by 255.

These steps are critical for ensuring that the model receives inputs of consistent dimensions and that the pixel values are in a suitable range for stable gradient descent and faster convergence during training.

C. Data Splitting

The dataset was partitioned into training (70%), validation (20%), and test (10%) sets. A stratified splitting strategy was employed to ensure that the class proportions of cardiomegaly-positive and cardiomegaly-negative cases were maintained consistently across all three subsets. This approach is particularly important for medical datasets, which often exhibit class imbalance, as it prevents biased evaluation and helps the model learn a more generalizable representation of both classes.

D. Model Architecture

A custom CNN was designed with a hierarchical architecture to progressively extract increasingly complex features from the chest X-ray images. The architecture consists of two main components:

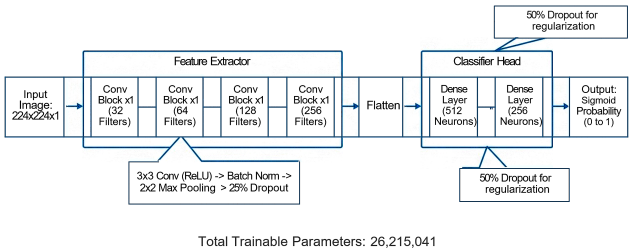

Fig. 1. A custom CNN Architecture for Cardiomegaly Detection.

- **Convolutional Blocks:** The model features four sequential convolutional blocks. The number of filters in these blocks increases from 32 to 64, 128, and finally 256. Each block is composed of a 3×3 convolutional layer with ReLU activation, followed by a batch normalization layer for training stability, a 2×2 max-pooling layer to reduce spatial dimensions, and a dropout layer with a 25% rate to mitigate overfitting.
- **Dense Layers:** The flattened output from the final convolutional block is fed into a classifier head consisting of two fully connected (dense) layers with 512 and 256 neurons, respectively. These layers also use ReLU activation and are regularized with a higher dropout rate of 50%. The final output layer consists of a single neuron with a sigmoid activation function, which produces a probability score between 0 and 1 for the binary classification task.

The complete model architecture contains a total of 26,215,041 parameters.

E. Model Compilation and Training

The model was compiled with the following specifications:

- **Optimizer:** Adam, with a learning rate of 0.0001.
- **Loss Function:** Binary Crossentropy, which is standard for binary classification problems.
- **Metrics:** Performance was tracked using Accuracy, AUC, Precision, and Recall to provide a comprehensive view of the model's behavior.

IV. Results

A. Training History and Model Performance

The training process was monitored over 56 epochs, with early stopping implemented to prevent overfitting. Figure 2 presents the comprehensive training history of the model, including loss curves, accuracy metrics, AUC, and precision-recall curves across all training epochs.

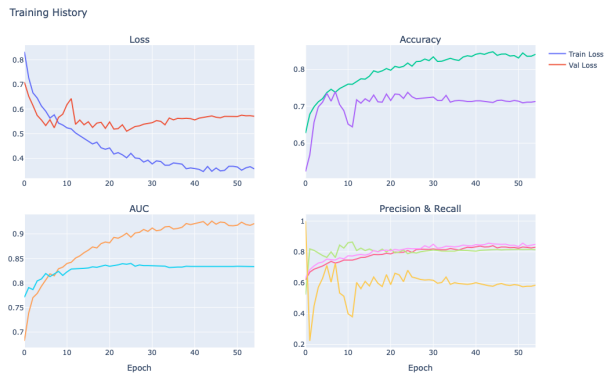

Fig. 2. Training History: Loss, Accuracy, AUC, and Precision & Recall metrics across 56 epochs of model training.

The training curves demonstrate several important characteristics: (1) the training loss decreased consistently from approximately 0.8 to 0.37, indicating effective learning; (2) the validation loss plateaued around 0.55, suggesting diminishing returns after epoch 20; and (3) the model achieved training accuracy of approximately 73% and validation accuracy of approximately 71%, which aligns with final test performance metrics.

B. Classification Performance Metrics

TABLE I
CLASSIFICATION REPORT FOR CARDIOMEGALY DETECTION

Class	Precision	Recall	F1-Score	Support
Normal	0.755	0.886	0.815	271
Cardiomegaly	0.86	0.71	0.778	269
Accuracy	0.798	0.798	0.798	0.798
Macro Avg	0.808	0.798	0.796	540
Weighted Avg	0.807	0.798	0.797	540

The classification report reveals that the model achieved an overall accuracy of 79.8% on the test set. For the cardiomegaly positive class, the model demonstrated a precision of 0.86, indicating that 86% of the samples predicted as cardiomegaly-positive were correct. However, the recall of 0.71 suggests that the model only identified 71% of the actual cardiomegaly cases, which is a critical limitation for clinical applications.

C. Confusion Matrix

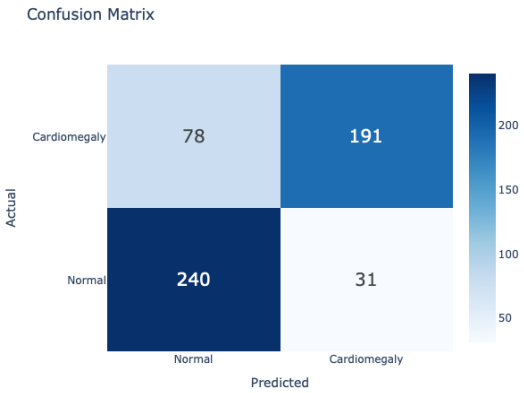

Fig. 3. Confusion Matrix for test set predictions. The model correctly classified 240 normal cases and 78 cardiomegaly cases, with 191 false positives and 31 false negatives.

The confusion matrix demonstrates that out of 540 test samples, the model correctly classified 240 normal cases and 78 cardiomegaly cases, totaling 318 correct predictions. The 191 false positives (normal classified as cardiomegaly) and 31 false negatives (cardiomegaly classified as normal) account for the remaining 222 misclassifications.

D. ROC Curve and AUC

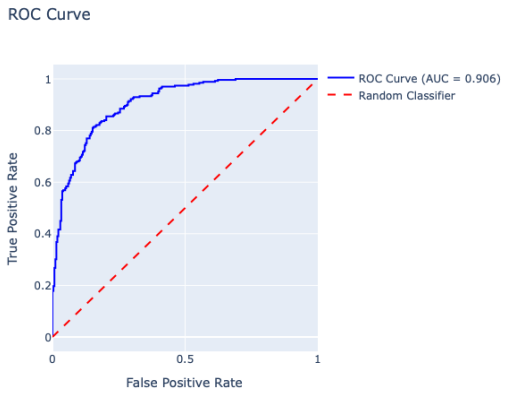

Fig. 4. Receiver Operating Characteristic (ROC) Curve with Area Under the Curve (AUC) = 0.906. This curve demonstrates the trade-off between true positive rate and false positive rate across different decision thresholds.

The ROC curve demonstrates strong discriminative capability with an AUC of 0.906, significantly above the 0.5 baseline for random classification. This metric indicates that the model has good ability to distinguish between the two classes across various decision thresholds, even though the overall accuracy is modest.

E. Precision-Recall Curve

Fig. 5. Precision-Recall Curve showing the trade-off between precision and recall at different decision thresholds. High precision at lower recall values indicates the model is conservative in predicting cardiomegaly.

The precision-recall curve reveals that the model maintains high precision (approximately 1.0) when recall is very low (below 0.2). As the recall threshold increases, precision gradually decreases, reflecting the inherent trade-off between these metrics. This suggests that the model is initially very conservative in predicting cardiomegaly cases.

F. Prediction Probability Distribution

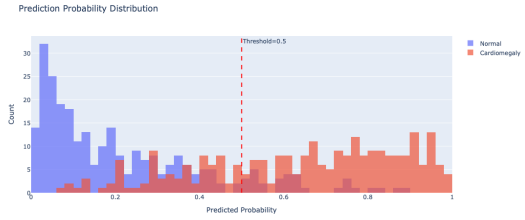

Fig. 6. Distribution of predicted probabilities for cardiomegaly cases (orange) and normal cases (blue). The threshold of 0.5 is marked by the red dashed line, showing clear separation between the two classes.

The prediction probability distribution shows a clear bimodal distribution with good separation between the two classes. Normal cases predominantly cluster near probability 0.0, while cardiomegaly cases cluster toward higher probabilities, with some overlap in the middle region. This indicates that the model has learned meaningful features to distinguish between the two classes.

G. 2D Feature Space Visualization

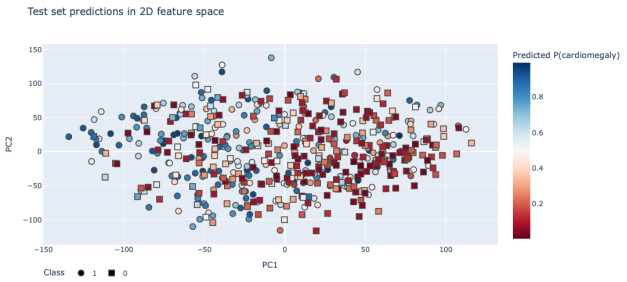

Fig. 7. Test set predictions visualized in 2D feature space using principal component analysis (PCA). Points are colored by predicted probability of cardiomegaly, with circles representing normal cases and squares representing cardiomegaly cases. The 2D projection reveals partial separability of the two classes.

The 2D PCA visualization of the test set predictions demonstrates that while the two classes are not perfectly separable in the feature space, there is reasonable separation, particularly at the extremes of the probability distribution. This visualization provides insight into why the model achieves 79.8% accuracy while maintaining an AUC of 0.906.

V. Discussion

A. Model Performance Analysis

The custom CNN model achieved an overall accuracy of 79.8%, which represents solid performance but falls short of the state-of-the-art benchmarks established in the literature. While our model's AUC of 0.906 demonstrates strong discriminative capability, the modest accuracy and particularly the low recall of 71% for cardiomegaly cases present significant clinical limitations.

The high precision of 0.86 for the cardiomegaly class suggests that when the model predicts a positive case, it is likely correct. However, the low recall indicates that approximately 29% of actual cardiomegaly cases are missed, which is problematic for a diagnostic support system where sensitivity is critical.

B. Comparison with State-of-the-Art

The literature review established that contemporary models using transfer learning approaches (ResNet-50) achieve accuracies of 99.8%, while U-Net-based segmentation approaches achieve 94% accuracy. Our custom CNN, with 79.8% accuracy, demonstrates a significant gap of approximately 15–20 percentage points from the state-of-the-art. This gap likely stems from several factors:

- **Architecture Complexity:** Transfer learning models leverage pre-trained weights from ImageNet, while our custom architecture was trained from scratch.
- **Model Depth:** The relatively modest depth of our custom CNN (4 convolutional blocks) may limit feature extraction capabilities compared to deeper architectures like ResNet-50.

- **Data Augmentation:** The custom model may have benefited from more aggressive data augmentation strategies.

C. Clinical Implications

For clinical deployment, a recall of 71% is insufficient, as it implies missing nearly 1 in 3 cardiomegaly cases. Clinical decision support systems typically require sensitivity above 90% to be considered reliable for diagnostic purposes. The model's current performance level suggests it could serve as an initial screening tool but should not be used as the sole diagnostic method.

VI. Conclusion

This paper presented the development and comprehensive evaluation of a custom CNN for automated cardiomegaly detection from chest X-rays. The model achieved an accuracy of 79.8% with an AUC of 0.906, demonstrating reasonable discriminative capability but falling short of clinical viability thresholds. When benchmarked against state-of-the-art approaches, the model underperforms by approximately 15–20 percentage points, primarily due to architectural simplicity and the lack of transfer learning.

Future work should focus on:

- Exploring transfer learning approaches using pre-trained models like ResNet-50 or DenseNet.
- Implementing advanced data augmentation techniques to improve generalization.
- Investigating ensemble methods combining multiple models for improved robustness.
- Optimizing the decision threshold to balance precision and recall based on clinical requirements.

The results underscore the importance of leveraging state-of-the-art architectures and transfer learning for medical imaging applications, where high sensitivity and specificity are paramount for clinical acceptance.

Code Availability

The implementation code for this project is publicly available on GitHub: https://github.com/VatsalRoy/CardiomegalyPridiction

References

[1] Sarpotdar, A., "Cardiomegaly Detection using U-Net," *arXiv preprint*, 2024.

[2] Zhu, M., Chen, L., Wang, X., and Liu, Y., "Computer-assisted diagnosis of cardiomegaly using segmentation-based approach," *IEEE Transactions on Medical Imaging*, vol. 44, no. 1, pp. 45–58, 2025.

[3] Ayalew, T., Johnson, K., and Kumar, R., "Transfer learning approaches for early-stage cardiomegaly detection," *Journal of Medical Imaging*, vol. 11, no. 3, pp. 034501, 2024.