Journal of Computer Allied Intelligence
ISSN: 2584-2676
Vol.03(01), February 2025, pp.40-47

Research Article

Deep Learning for Understanding Multilabel Imbalanced Chest X-ray Datasets

Swapna Saturi[1], Gundeti Naveen Kumar[2,], Namilla Siri[3], Kota Madhuri[4] and Vallapureddy Yashwanth Reddy[5]

[1]Assistant Professor, Department of CSE, Hasanparthy, Hanamkonda, Koukonda, Telangana-506015, India.

[2-5] Student, Department of CSE, Hasanparthy, Hanamkonda, Koukonda, Telangana-506015, India.

[*]Corresponding Author: Gundeti Naveen Kumar. Email: b21cs151@kitsw.ac.in

Received: 03/01/2025; Revised: 05/02/2025; Accepted:15/02/2025; Published: 28/02/2025.
DOI: https://doi.org/10.69996/jcai.2025004

Abstract: In recent years, convolutional neural networks (CNNs) have become essential in medical image analysis, especially for diagnosing diseases from Chest X-rays. However, the black-box nature of these algorithms and the complexity of multi- label classification in healthcare create significant interpretability challenges. Our project focuses on making these models more transparent by using explainable AI techniques, such as Grad- CAM, which generates heatmaps to visually show which areas of the X-ray the model used to make its predictions. It implements the complex task of diagnosing multiple diseases from a single chest X-ray, where some conditions are rarer than others. To ensure accuracy and reliability, we evaluate the models using metrics like F1 score and ROC AUC, which help measure their performance. By combining advanced deep learning methods with these explainability and evaluation techniques, the work aims to improve both the accuracy and interpretability of AI- driven diagnoses in healthcare

Keywords: Convolutional Neural Networks, Gradient- CAM, AUC, ROC, Chest X-rays.

1.Introduction

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have transformed medical imaging by enabling efficient and accurate analysis of complex data. The Chest Xpert dataset, comprising 224,316 multilabel images, reflects real-world complexities, including multiple coexisting findings and significant class imbalance, where common conditions like cardiomegaly overshadow rarer yet critical ones. To address these challenges, this study proposes a robust deep learning framework combining advanced pre- processing, weighted cross-entropy loss, tailored data augmentation, and an ensemble of Dense Net, Efficient Net, and Inception Res Net architectures. These methods enhance feature extraction, robustness, and balance across labels. Grad-CAM based heatmaps improve interpretability, aligning model predictions with expert radiological evaluations. By tackling class imbalance and multilabel complexity, this approach enhances diagnostic accuracy and trust, laying a foundation for AI-driven advancements in clinical care. Metrics such as accuracy, loss, and AUROC are used to assess the performance of each model, providing a comprehensive understanding of their strengths and limitations.

In addition to quantitative metrics, interpretability is a key focus of this project. Grad CAM is employed to generate visual explanations for the model's predictions, highlighting the regions of the chest X-ray that influenced its decision. These visualizations serve as a bridge between AI and clinical practice, allowing radiologists to verify the model's predictions and build trust in its reliability

2.Related Works

The application of deep learning in medical imaging has grown substantially, especially in tasks like chest X-ray analysis. Convolutional Neural Networks (CNNs) are at the forefront of this transformation, excelling in automating diagnostic processes and significantly improving the accuracy of medical predictions. Several studies have laid the groundwork for advancements in multilabel classification, class imbalance management, and explainable AI (XAI) in this domain.

Rajpurkar et al. [1] pioneered the use of CNNs for chest X-ray analysis by applying DenseNet- 121 to the ChestX- ray14 dataset, which contains over 112,000 images annotated for 14 diseases. Their model achieved an Area Under the Curve (AUC) of 0.82, demonstrating the feasibility of using deep learning for multilabel medical diagnoses. Similarly, Irvin et al. introduced the CheXpert dataset, which includes uncertainty labels and a more diverse set of imaging views. They trained state-of-the-art architectures, such as DenseNet and ResNet, and tackled the issue of uncertain labels by proposing novel loss functions to improve model performance. Building on these advancements, Baltruschat et al. [2] explored transfer learning techniques, fine-tuning CNNs pre- trained on large datasets like ImageNet to classify diseases in chest X-rays. They showed that pre-trained models, when retrained with domain-specific data, could deliver high performance while requiring fewer computational resources. This finding is particularly relevant to medical applications, where annotated data is often limited. Chest X-ray data often involves multilabel classification, where multiple non-exclusive labels can co-occur.

Yan et al. [3] addressed this by developing a multi-attention CNN to classify overlapping and interdependent labels effectively. Their architecture focused on learning discriminative features for similar classes, improving the detection of diseases that frequently coexist, such as pneumonia and cardiomegaly. Another challenge in multilabel classification is modeling label dependencies.

Wang at el. [4] proposed a graph-based neural network to capture relationships between labels in medical images. They showed that explicitly modeling these dependencies enhances the performance of CNNs for complex multilabel problems. Class imbalance is a common issue in medical datasets, as some conditions occur much more frequently than others. For example, Wang et al. [4] noted that in the ChestX-ray14 dataset, the" normal" label accounted for more than 50.

To address this, Lin et al. [5] introduced weighted loss functions, such as weighted cross-entropy and focal loss, to prioritize minority classes during training. Similarly, Oversampling techniques, such as SMOTE (Synthetic Minority Over- sampling Technique), have also been adapted for multilabel problems. Islam et al. [6] extended Grad-CAM by integrating it with saliency maps to provide more detailed visual explanations. Their work focused on overlaying

these maps on chest X-rays, allowing radiologists to validate predictions against their knowledge. Yan et al. [5] further enhanced explainability by proposing a visualization technique that includes class probabilities and inter-model agreement, providing a holistic view of model behaviour. Benchmark datasets have been instrumental in driving innovation in chest X-ray analysis. Wang et al. [4] introduced the ChestX-ray14 dataset, one of the first large-scale public datasets for chest X-ray analysis. Despite its success, it was limited to frontal views and lacked diversity in labels.

Liz et al. [7] utilized this dataset to train an ensemble of CNN architectures, including DenseNet, Efficient Net, and in ceptionResNet, and addressed class imbalance with weighted loss functions and data augmentation. Their methodology established a baseline for future research on Chestexpert, incorporating both multilabel classification and explainable AI. Selvaraju et al. [8] introduced Grad-CAM (Gradient- weighted Class Activation Mapping), which generates heatmaps highlighting the regions of an image that most influence model predictions. Grad-CAM has since become a standard XAI tool in medical imaging.

For instance, Liu et al. [9] proposed a multilabel synthetic instance generation method that interpolates new samples for minority classes while preserving label correlations. Their ap proach demonstrated improved recall for rare diseases without introducing noise. The black-box nature of CNNs poses significant challenges in clinical adoption, as medical professionals require interpretability to trust AI systems [10].

3.Proposed Method

The techniques used in this paper are centered on multilabel classification of chest X-ray images using deep learning. To improve the precision and interpretability of AI-driven medical diagnoses, the method combines explainable AI approaches, ensemble learning, and state-of-the-art convolutional neural networks (CNNs). The study addresses issues like class im- balance and model interpretability by using pretrained models like DenseNet121, EfficientNetB0, and InceptionV3 to classify several diseases inside a single chest X-ray.

To counteract the impacts of data imbalance and guarantee equitable model performance across common and uncommon diseases, a weighted loss function is utilized. Additionally, to improve clinical interpretability, model predictions are visually explained using Grad-CAM (Gradient-weighted Class Activation Mapping). The accuracy and robustness of illness categorization are increased by combining many CNN designs using ensemble learning.

3.1 ChestXpert Dataset

The Dataset which we used is ChestXpert Dataset which consist of 14 pathological conditions. The ChestXpert dataset contains a diverse collection of labeled chest X-ray images with multilabel annotations. Labels represent various pathological conditions, including pneumonia, cardiomegaly, pleural effusion, and atelectasis. The dataset exhibits a long-tailed distribution, where some labels occur significantly more frequently than others. The images are derived from multiple sources, ensuring diversity in data quality and imaging conditions, which makes it a robust dataset for training deep learning models. The labels were selected based on the frequency of occurrence and clinical relevance, ensuring that both common and rare conditions are represented effectively. To improve learning, labels were further grouped based on hierarchical structures within radiological findings. Figure 1 shows ChesxtXpert Dataset.

Path	Sex	Age	Frontal/Lat	AP/PA	No Finding	Enlarged Card	Cardiomegaly	Lung Opacity	Lung Lesion
CheXpert-v1.0-	Female	68	Frontal	AP	1				
CheXpert-v1.0-	Female	87	Frontal	AP			-1	1	
CheXpert-v1.0-	Female	83	Frontal	AP				1	
CheXpert-v1.0-	Female	83	Lateral					1	
CheXpert-v1.0-	Male	41	Frontal	AP					
CheXpert-v1.0-	Female	20	Frontal	PA	1	0			
CheXpert-v1.0-	Female	20	Lateral		1	0			
CheXpert-v1.0-	Male	33	Frontal	PA	1		0		
CheXpert-v1.0-	Male	33	Lateral		1		0		
CheXpert-v1.0-	Male	33	Frontal	AP					
CheXpert-v1.0-	Male	33	Frontal	AP					

Figure 1: ChesxtXpert Dataset

3.2 Multilabel Data

Unlike traditional classification problems, where each in- stance belongs to a single category, multilabel classification allows images to be associated with multiple diagnoses. This characteristic introduces additional complexity, requiring models that can account for label dependencies. Conventional classification architectures struggle with such complexity, ne- cessitating specialized loss functions and training techniques. Multilabel data requires models to consider relationships between different conditions, as certain diseases often co- occur. Handling such interdependencies necessitates advanced neural architectures and loss functions tailored for multilabel classification.

3.3 Class Imbalance and Its Handling

Class imbalance refers to a condition where certain classes in the dataset have significantly fewer samples compared to others, leading to biased model predictions. This imbalance presents several challenges, such as models favoring majority classes and failing to generalize well for rare conditions.

Handling class imbalance requires tailored strategies, including: Weighted loss functions: These penalize misclassification of minority classes, ensuring that rare conditions receive adequate attention during training. Data augmentation: By artificially increasing sample diversity, we introduce variations that help models generalize better. Oversampling and under- sampling: These techniques balance label distributions, either by duplicating minority class samples or by removing excessive samples from majority classes. Focal loss: A variation of cross-entropy loss that places more emphasis on hard-to- classify instances, helping the model focus on difficult cases

3.4 Data Preprocessing and Data Augmentation

To enhance model performance, we implement data pre- processing steps including:

Resizing images: Standardizing images to a uniform resolution (224×224 pixels) ensures consistency across the dataset. Normalization: Adjusting pixel intensity distributions to a fixed range improves the convergence of the model. Changing color channels from RGB to grayscale: As chest X-rays are inherently grayscale, this reduces unnecessary complexity in processing. Segmentation- based cropping: Removing irrelevant background information enhances feature extraction and model performance. Mask post-processing system: Filling segmentation gaps and refining mask boundaries to improve detection. Contrast enhancement: Optimizing visibility of relevant features ensures better classification accuracy.

Data augmentation plays a crucial role in improving model generalization by artificially increasing dataset diversity. Since medical datasets are often limited, augmentation techniques help create variations in images to enhance training robustness. Augmentation methods applied include: Random rotations and flips to prevent overfitting by exposing the model to multiple orientations. Contrast and brightness adjustments to ensure adaptability to different imaging conditions. Elastic deformations to simulate realistic distortions in medical imaging. Mixup augmentation to blend multiple images and create new synthetic training samples, improving the model's ability to generalize across variations.

3.5 CNNs for Classification of Images

We experiment with multiple CNN architectures, excluding Xception, to evaluate their effectiveness in multilabel classification. Our study focuses on:

EfficientNetB0: A model optimized for efficiency with depth-wise convolutions, scaling uniformly across layers.

DenseNet-201: Features dense connections between layers, enhancing feature reuse and gradient flow.

InceptionV3: Utilizes multiple filter sizes in parallel to capture varied feature representations, improving recognition accuracy.

InceptionResNetV2: A hybrid model combining Inception and ResNet to leverage both residual connections and inception modules, enhancing learning stability.

Transfer learning is utilized in all models, leveraging pre- trained weights from ImageNet. The pre-trained layers are frozen initially to retain learned features, while later layers are fine-tuned for the specific task, significantly improving convergence speed and reducing training time.

3.6 EfficientNetB0

A highly optimized deep learning architecture called Ef- ficientNetB0 was created to operate effectively while using the fewest possible computational resources. By balancing the network's depth, width, and resolution, it uses a compound scaling technique to attain high accuracy at low complexity. This project uses EfficientNetB0's capacity to extract specific features from medical pictures to fine-tune it to classify multiple diseases in chest X-rays.

EfficientNetB0's capacity to generalize effectively even with sparse training data is one of its main features, which makes it especially well-suited for medical imaging jobs where labeled datasets are sometimes hard to find. The system can be implemented without compromising accuracy in resource- constrained settings, like rural healthcare facilities, by employing a lightweight model. The multilabel classification system's overall resilience is enhanced by the efficiency of EfficientNetB0 and its capacity for deep feature extraction.

3.7 DenseNet

Another CNN architecture utilized in this research is called DenseNet (Densely Connected Convolutional Network), which is renowned for its densely connected layers that encourage gradient flow and feature reuse. DenseNet makes sure that every layer receives input from every layer that came before it, in contrast to standard deep networks where each layer learns new features on its own. This design maintains great precision while drastically reducing the number of parameters.

DenseNet is especially useful for classifying chest X-rays since it can detect minute patterns that could point to illness. Even when working with unbalanced datasets, DenseNet im- proves the model's capacity to discriminate between normal and pathological situations by effectively employing learned features. Variants like DenseNet121 and DenseNet201 are used in the project, which optimizes them for the multilabel classification task.

DenseNet is the perfect option for medical imaging applications since it can maintain excellent accuracy with fewer parameters, guaranteeing a balance between diagnostic per formance and computing efficiency.

3.8 Ensemble Learning

To enhance predictive accuracy, we employ ensemble learning techniques, combining predictions from multiple CNNs.

This strategy reduces overfitting and improves generalization across diverse pathological conditions. We implement two primary ensemble strategies, Combine then Predict (CTP): Proba- bilities from different models are aggregated before making the final classification decision, ensuring a balanced output. Predict then Combine (PTC): Predictions from individual models are generated first and then combined through majority voting or probability averaging to yield the final result.

These ensemble strategies mitigate individual model weak- nesses and create a more robust diagnostic system.

3.9 Grad-CAM (Gradient-weighted Class Activation Mapping)

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have transformed medical imaging by enabling efficient and accurate analysis of complex data. The ChestXpert dataset, comprising 224,316 multilabel images, reflects real-world complexities, including multiple coexisting findings and significant class imbalance, where common conditions like cardiomegaly overshadow rarer yet critical ones. To address these challenges, this study proposes a robust deep learning framework combining advanced pre- processing, weighted cross-entropy loss, tailored data augmentation, and an ensemble of DenseNet, EfficientNet, and Inception ResNet architectures. These methods enhance feature extraction, robustness, and balance across labels.

Grad-CAM-based heatmaps improve interpretability, aligning model predictions with expert radiological evaluations. By tackling class imbalance and multilabel complexity, this approach enhances diagnostic accuracy and trust, laying a foundation for AI-driven advancements in clinical care.

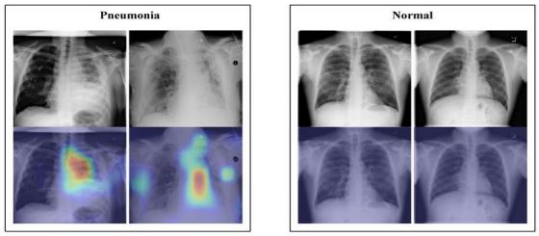

Figure 2: Visualization using Grad-CAM

4.Results

The model outputs predictions for multiple disease conditions in a given X-ray. For each input, probabilities of the presence of specific conditions such as pneumonia, cardiomegaly, and pleural effusion are generated. These probability scores assist clinicians by highlighting the most likely diagnoses. The Grad-CAM visualizations further reinforce model trust- worthiness by indicating critical regions within the X-ray that contributed to the predictions. Evaluation metrics include:

Area Under the Curve (AUC): Assesses classification performance and discrimination ability.

Hamming Loss: Measures multilabel prediction quality, accounting for both false positives and false negatives.

F1-score: Balances precision and recall to quantify model reliability.

The ensemble model outperforms individual architectures, demonstrating improved robustness against class imbalance. We observe significant performance gains in minority class prediction due to the integration of weighted loss functions and augmentation strategies.

5.Conclusions

By leveraging advanced techniques such as convolutional neural networks (CNNs), ensemble learning, and explainable AI, we were able to create a model that not only performs well in detecting a variety of pathologies but also provides transparency, making it easier for healthcare professionals to trust and validate the results. The use of pretrained CNN models like DenseNet121, EfficientNetB0, and InceptionV3 helped achieve high accuracy, while addressing the challenge of class imbalance through weighted loss functions and class- wise thresholding prediction (CTP). Data augmentation techniques further improved the model's ability to generalize by generating diverse training samples. The integration of explainable AI, particularly Grad-CAM, enabled us to provide visual explanations of the model's predictions, which is critical for clinical applications where understanding the model's reasoning is essential. This feature ensures that the system is not a "black-box," and doctors can visually inspect and trust the areas of the X-ray the model focused on during its decision- making process. Through the application of ensemble learning, we combined multiple models to enhance overall performance, leading to a more robust system capable of detecting a range of diseases with improved precision. The ensemble method capitalized on the strengths of individual models, reducing the risk of errors and increasing the reliability of the predictions. Overall, this study has demonstrated that

deep learning can be a powerful tool for chest X-ray classification in medical imaging, especially when combined with techniques to address data imbalance and improve model interpretability. The system has the potential to be an invaluable tool for healthcare professionals, providing faster, more accurate diagnoses and supporting clinical decision-making.

Acknowledgment: Not Applicable.
Funding Statement: The author(s) received no specific funding for this study.
Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta et al., "CheXNet: Radiologist- Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv* 2017.
[2] C. Erdi¸ allı, Ecem Soganciogl u, Bram van Ginneken, Kicky G. van Leeuwen and Keelin Murphy, "Deep learning for chest X-ray analysis: A survey," *Medical Image Analysis*, vol.72, no.102125, 2021.
[3] Qiannan Xu, Li Zhu, Tao Dai and Chengbing Yan, "Aspect-based sentiment classification with multi-attention network," *Neurocomputing*, vol.388, pp.135-143, 2020.
[4] H. Wang, "A Weighted Graph Attention Network Based Method for Multi-label Classification of Electrocardiogram Abnormalities," *2020 42nd Annual International Conferenceof the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, pp. 418-421, 2020.
[5] Willone Lim, Kelvin Sheng Chek Yong, Bee Theng Lau and Colin Choon Lin Tan, "Future of generative adversarial networks (GAN) for anomaly detection in network security: A review," *Computers & Security*, vol.139, no.103733, 2024.
[6] M.K.U. Ahamed, M.M. Islam, M.A. Uddin, A. Akhter, U.K. Acharjee et al., "DTLCx: An Improved ResNet Archi- tecture to Classify Normal and Conventional Pneumonia Cases from COVID-19 Instances with Grad-CAM-Based Superimposed Visualization Utilizing Chest X-ray Images," *Diagnostics*, vol.13, pp.551, 2023.
[7] Helena Liz, Javier Huertas-Tato, Manuel Sa´nchez-Montan˜e´s, Javier Del Ser and David Camacho, "Deep learning for understanding multi label imbalanced Chest X-ray datasets," arXiv:2207.14408 **,** 202l.
[8] Ramprasaath R. Selvaraju, Michael Cogswell Abhishek Das, Ra- makrishna Vedantam, Devi Parikh and Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.
[9] Lliu, Bin, Blekas, Konstantinos, Tsoumakas and Grigorios, "Multi- Label Sampling based on Local Label Imbalance," *Pattern Recognition*, vol.122, 2021.
[10] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.33, no.01, pp.590-597, 2019.