CENG 407

Innovative System Design and Development I

2025-2026 Fall

MULTI-LABEL CHEST X-RAY DISEASE CLASSIFICATION AND EXPLAINABILITY WITH DEEP LEARNING

LITERATURE REVIEW

YAREN AKDOĞAN 202111046

SEZİN KÖROĞLU 202211046

İBRAHİM TAŞPINAR 202111072

CEVDET ALP ÖZGÜN 202211406

İPEK BAŞ 202211013

SALİH BARKIN AKKAYA 202211004

Abstract

In today's medical field, chest X-ray images have an indispensable role in the diagnosis of thoracic diseases such as pneumonia, cardiomegaly and atelectasis. Quick and accurate diagnosis is crucial but the growing number of patients and the limited availability of radiologists can delay the results of X-rays. Deep learning techniques such as convolutional neural networks have shown great potential in the automation of disease detection thanks to their ability to detect multiple diseases and them allowing for models with accuracies comparable to human professionals. However, high accuracy is not the only relevant factor for real world clinical adoption. Doctors need to understand why a model makes a certain decision in order to trust and effectively use it. This is where explainable artificial intelligence methods such as Grad-CAM play an important role by visually highlighting the parts of an image that are relevant to the model's prediction through heatmaps. Through this review, we aim to understand the challenges we may face in development and methods we can employ in overcoming them. We also attempt to understand methods used previously in studies to improve the interpretability and accuracy of models for clinical us

Table of Contents

1. Introduction……………………………………………………………………..1

2. Deep Learning and Convolutional Neural Networks…………………………………1

2.1. Fundamentals of Convolutional Neural Networks………………………………1

2.2. Transfer Learning in CNN Based Medical Imaging……………………………...1

2.3. CNN Architectures Used in Chest X-Ray Analysis………………………………2

2.3.1. AlexNet…………………………………………………………………2

2.3.2. ResNet & DenseNet……………………………………………………2

2.3.3. Transformer (SwinCheX)……………………………………………....2

3. Datasets and Multi-Label Classification………………………………………………2

3.1. NIH ChestX-ray14………………………………………………………………. 2

3.2. CheXpert…………………………………………………………………………2

3.3. Multi-Label Classification Challenges……………………………………………3

4. Explainable AI (XAI) Approaches……………………………………………………3

4.1. Importance of Explainability……………………………………………………..3

4.2. Grad-CAM……………………………………………………………………….3

5. Findings from Related Works…………………………………………………………4

5.1. DenseNet121, CheXNet, and Other Chest X-Ray Models………………………4

5.2. Uncertainty and Class Imbalance………………………………………………...4

5.3. Explainability and Grad-CAM…………………………………………………...5

5.4. Limitations of Current Models and Multimodal Models………………………...5

5.5. Summary…………………………………………………………………………5

6. Conclusion……………………………………………………………………………6

7. References……………………………………………………………………………7

1. Introduction

Chest X-ray images play an important role in diagnosing thoracic conditions, including pneumonia, cardiomegaly, and atelectasis [1,4]. The increasing volume of medical images, along with limited expert availability, has motivated the development of automated diagnosis systems based on deep learning [2]. Multi-label classification is particularly challenging in this context, as a single chest X-ray image may indicate multiple illnesses. Convolutional neural networks (CNNs) have become the dominant approach for image based medical diagnosis [2]. Transfer learning from large scale datasets such as ImageNet has further improved performance gains by providing pre-trained feature extractors. Alongside high classification accuracy, explainability has become one of the most important requirements for adoption in hospitals. Explainable AI (XAI) methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM), allow doctors to view which parts of an image influence a model's prediction [4]. This literature review focuses on deep learning models for multi-label chest X-ray classification, dataset characteristics, and explainability methods, in an attempt to understand what is required to improve on existing ideas and possible difficulties we may encounter during development.

2. Deep Learning and Convolutional Neural Networks

2.1. Fundamentals of Convolutional Neural Networks

Convolutional Neural Networks are a type of deep learning model particularly suited for image classification due to their ability to extract features via convolutional layers. Convolutional layers apply filters to the input image to learn patterns such as edges, textures, or shapes. As data goes through layers, the model learns how to recognize objects such as organs or lesions in medical images. Pooling layers are used to reduce the size of feature maps to decrease the computational cost and to allow the model to only focus on the most significant information. Activation functions are used to add non-linearity to models to allow models to learn more complex patterns. By applying these layers several times, models learn to differentiate and classify images effectively.

2.2. Transfer Learning in CNN Based Medical Imaging

Due to the limited availability of medical image datasets, transfer learning is widely adopted in an effort to make the most of the lacking data. CNNs that have been pre-trained on large datasets to detect general objects are fine-tuned on chest X-ray images [2,7]. This method allows models to benefit from the detection capabilities of the pre-trained models, while still being customizable enough for optimization for medical images. This approach also decreases the required training time for models as less work needs to be done during feature extraction compared to a standard CNN model. Although there are many different pre-trained model alternatives with different capabilities, ResNet, AlexNet and DenseNet appear to be the most popular [1,2,4,5,6,7].

2.3. CNN Architectures Used in Chest X-Ray Analysis

2.3.1. AlexNet

AlexNet has been applied in chest X-ray classification for pneumonia and other thoracic diseases. For example, [7] used AlexNet for pneumonia detection, reporting around 81% accuracy. In another study, [5] combined AlexNet features with image processing methods to classify multiple thoracic conditions.

2.3.2. ResNet & DenseNet

More modern CNNs, such as ResNet and DenseNet, are widely used due to better feature extraction and deeper architectures. Comparative studies on tuberculosis and COVID-19 detection show that DenseNet121 and ResNet variants achieve high AUC scores and outperform AlexNet in multi-label chest X-ray classification [1,2,3,4,6].

2.3.3. Transformer (SwinCheX)

Although CNNs dominate, transformer-based architectures have begun to emerge. SwinCheX [6] applied a Swin Transformer backbone to the ChestX-ray14 dataset and achieved an average AUC of 0.810, demonstrating that transformer architectures can also perform competitively in medical imaging tasks.

3. Datasets and Multi-Label Classification

3.1. NIH ChestX-ray14

The NIH ChestX-ray14 dataset contains over 100,000 frontal chest X-ray images labelled for 14 different thoracic diseases, and it has become one of the most commonly used datasets in this research field [1]. Because manually labelling medical images is slow and requires expert radiologists, the dataset uses NLP methods to automatically extract labels from the original radiology reports. This approach makes it possible to build a very large dataset, but it also means that some labels may be inaccurate or wrong due to the uncertainty tied to automated text processing [1]. Even with this drawback, ChestX-ray14 is the most popular choice for studies thanks to its size and easy accessibility for researchers [1].

3.2. CheXpert

CheXpert is a large chest X-ray dataset with over 224,000 images from 65,000 patients [4]. Although the higher image quantity is a positive, the main feature that differentiates this dataset from the NIH ChestX-ray14 dataset is that CheXpert includes uncertainty labels, which capture cases where the radiology reports are ambiguous or unclear. Properly handling these uncertain labels is important for training models that achieve high accuracy.

Researchers often use strategies such as U-Ones, U-Zeros, or U-Ignore, depending on the specific disease being predicted, to make the most of the dataset while managing uncertainty [4]. The reason this dataset has seen less use is most likely due to it being newer compared to the NIH ChestX-ray14 dataset and its harder access requirements. This dataset has started overtaking the NIH ChestX-ray14 dataset in popularity in recent years due to it providing better results for most studies.

3.3. Multi-Label Classification Challenges

Unlike single label classification, multi-label chest X-rays require predicting multiple diseases present at once [1,4]. In practice, most studies use binary cross-entropy based losses such as BCEWithLogitsLoss to allow the model to treat each disease as its own prediction. However, some diseases appear far less often than others, which makes training uneven. To deal with this, previous studies have often given more weight to rare classes or have adjusted the loss based on how many samples each class has [1,4]. Others simply increased the number of images for less common diseases through oversampling or augmentation [1,4]. These strategies help the model reach a higher final accuracy.

4. Explainable AI (XAI) Approaches

4.1. Importance of Explainability

In clinical applications, model interpretability is critical. Doctors need to understand why a model made a particular decision to trust its output [2,3,4]. XAI methods provide this capability by highlighting regions influencing model predictions.

4.2. Grad-CAM

Grad-CAM is an explainability method that identifies which regions of an image contribute most to a model's decision by calculating the gradient of a target class score with respect to the CNN. This produces a heatmap that highlights the most relevant areas for the prediction. Because it is simple to implement and can be added to existing CNN models without retraining, Grad-CAM has become one of the most commonly used tools for interpreting deep learning based medical imaging systems.

However, Grad-CAM comes with several limitations. Its visualizations provide only coarse localization and are not completely pixel accurate, which can be problematic for detecting small or subtle abnormalities [4]. Although for most diseases this is a minor issue, some diseases with subtler or smaller indications suffer due to this. Additionally, Grad-CAM shows what influenced the model rather than why the model was influenced, meaning the highlighted regions do not always represent the correct disease [4]. This can create uncertainty when doctors rely on the visualization to validate the model's decisions.

5. Findings from Related Works

5.1. DenseNet121, CheXNet, and Other Chest X-Ray Models

DenseNet121, particularly with the CheXNet dataset, continues to be widely adopted as a strong baseline architecture for multi-label chest X-ray classification [2,3,4,6]. CheXNet's densely connected layer architecture allows features from earlier layers to be reused in later layers preventing vanishing gradients and other similar problems which are common in deep learning.

CheXNet [2] demonstrated performance exceeding average radiologists in detecting pneumonia, achieving an AUROC of 0.76 to 0.84 across different test subsets. Subsequent studies using the NIH ChestX-ray14 dataset applied this approach to 14 thoracic diseases, reporting average AUROC values ranging from 0.80 to 0.87 depending on preprocessing and training methods [3,4]. In an attempt to further validate the findings of these studies, we have also trained a reimplementation of the original CheXNet model in Python 3. Our findings aligned almost exactly with the original study with an AUROC score of around 0.76 despite hardware limitations. These results indicate DenseNet121 as a good fit to be a backbone for chest X-ray predictions.

We have also attempted to train the transformer model SwinCheX used in an effort to compare transformer and CNN architecture results. The original SwinCheX study achieved a high average AUROC score of 0.81 [6]. Our retrained model experienced difficulties due to lacking hardware and as such did not utilize the optimal training parameters specified by the study. Due to this issue our model only reached an average AUC score of 0.54. From these training results we have deduced that the transformer architecture is not viable for lower end hardware.

5.2. Uncertainty and Class Imbalance

The currently widely available chest X-ray dataset all present a few common issues such as class imbalance due to rare or hard to find diseases and uncertain labels due to the utilization of NLPs in the automated labelling of the data. Many classes, such as pneumothorax or hernia, occur rarely compared to more common conditions like cardiomegaly or effusion. In the NIH ChestX-ray14 dataset this issue is quite noticeable as some findings only have around 300 labels whereas the more common ones can have up to 20000 instances [1].

In addition, automated label extraction from radiology reports introduces ambiguity and potential errors. For instance, ChestX-Ray8[1] found that weakly supervised labelling could mislabel up to 10% to 15% of cases due to uncertainty in the original text reports [1]. To address these issues, studies have employed methods such as weighted cross-entropy loss, focal loss, or class-balanced loss to give rarer conditions greater influence during training. U-Ignore and U-Ones strategies for handling uncertain labels in CheXpert have been proposed, demonstrating that appropriate treatment of uncertainty can improve AUROC by 2 to 3 percent on some diseases [4]. These findings indicate the importance of proper preprocessing and appropriate loss design to achieve consistent and accurate performance across all classes.

5.3. Explainability and Grad-CAM

While model accuracy is essential, clinical adoption requires interpretable predictions. Grad-CAM is frequently used to visualize which areas of an image influence a model's output [4]. CheXpert applied Grad-CAM to highlight lesions indicating pneumonia, showing that heatmaps generally overlapped with regions on the X-ray identified by radiologists [4].

However, Grad-CAM provides only rough localization, and pixel perfect interpretability has not yet been achieved in this area while using such tools. Later studies experimented with hybrid approaches where they combined Grad-CAM with attention mechanisms or integrated gradients to improve resolution [4,6]. These enhancements are particularly useful in finding subtle diseases, such as early stage edema or nodules, which are harder to find due to their small size, where precise localization can inform the clinical decisions of doctors.

5.4. Limitations of Current Models and Multimodal Models

Most existing models rely completely on frontal X-ray images, ignoring lateral views or complementary clinical data such as patient history, lab results, or prior imaging. Limiting the input to a single type of data decreases model performance. For example, several studies reported that including other patient data such as age, sex, and prior diagnoses improved average AUROC scores by 1 to 2 percent for certain rare diseases [4,6]. Multimodal models appear to be the next major step in a path toward models that are both more accurate and better aligned with the interests of doctors.

5.5. Summary

Overall, the literature shows DenseNet-121 based architectures as the most commonly used method for multi-label chest X-ray classification [2,3,4,6] and shows explainability methods like Grad-CAM as a critical requirement for trust from doctors [4]. Major challenges remain, including handling uncertain or wrong labels, class imbalance and integrating additional relevant patient information for multimodal predictions. Addressing these issues is important for the development of systems that are not only accurate but also interpretable, trustworthy and clinically relevant.

6. Conclusion

Deep learning has allowed multi-label chest X-ray classification to advance to a state where it's viable in the real world. CNNs such as DenseNet-121 achieve high accuracy on the NIH ChestX-ray14 and the CheXpert datasets. Explainable AI techniques such as Grad-CAM have allowed doctors to visually understand model predictions, improving the interpretability of the models. Difficulties remain in handling rare or subtle diseases, uncertain or wrong labels and incorporating new data sources that may be relevant to patients. Future research can focus on higher resolution XAI, and optimizing architectures to improve clinical utility. This literature review demonstrates that while there has been significant development in the area, more attention towards explainability and dataset quality is required for real world adoption of this new technology as a form of trusted medical diagnosis.

7. References

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

[2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225.* https://arxiv.org/abs/1711.05225

[3] Rajpurkar, P., Irvin, J., Bagul, A., Ko, M., Taylor, R., Duan, T., Poplin, R., Chen, P., Jones, E., Seal, P., & Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine, 15*(11), e1002686. https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.1002686&type=printable

[4] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Wilson, M., Mooney, C., Anand, V., Pursnani, D., Goyal, P., Naidoo, K., ... Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *AAAI Conference on Artificial Intelligence.* https://arxiv.org/abs/1901.07031

[5] Akmal, M. R., Bintoro, K. B., & Purboyo, W. T. (2019). Chest X-Ray Image Classification on Common Thorax Diseases using GLCM and AlexNet Deep Features. *International Journal of Integrated Engineering, 11*(4), 183–193.

[6] Taslimi, S., Taslimi, M., & Aghabozorgi, S. (2022). SwinCheX: Multi-label classification on chest X-ray images with transformers. *arXiv preprint arXiv:2206.04246.* https://arxiv.org/abs/2206.04246

[7] Gao, Y., Xie, H., Chen, R., Huang, Y., Guo, Y., Zhang, Y., & Dong, F. (2025). Pneumonia Detection and Analysis Using AlexNet. *Applied and Computational Engineering, 190*(1), 1–7. https://www.researchgate.net/publication/396128636_Pneumonia_Detection_and_Analysis_Using_AlexNet