



A Review on Deep Learning for Pulmonary Disease Classification Through CheXpert Dataset's Chest X-Ray Images

Kirti Swagat Mohanty¹, Chandrajit Choudhury², Prabina Pattanayak³

¹Department of Electronics and Communication Engineering, National Institute of Technology, Silchar, 788010, Assam, India. Email: kirti21rs@ece.nits.ac.in

²Department of Electronics and Communication Engineering, National Institute of Technology, Silchar, 788010, Assam, India. Email: chandrajit@ece.nits.ac.in

³Department of Electronics and Communication Engineering, Specialization: Wireless Communications, MIMO Communications, Massive MIMO Communications, Smart Grid Communications, Soft Computing, National Institute of Technology Warangal, 506004, Telangana, India. Email: prabina@nitw.ac.in

ABSTRACT

Chest radiography stands out as a common imaging modality for the screening and monitoring of pulmonary disease, but its interpretation is hindered by over-lapping presentations, image variability and inter-observer disagreement. Deep learning offers a path towards consistent, scalable computer-aided diagnosis, and public release of CheXpert dataset has catalysed a substantial body of method-ological work. This literature survey presents a systematic, PRISMA-guided review of deeplearning approaches developed and evaluated on CheXpert dataset for the multi-label classification of thoracic pathologies. From an initial pool of 197 records identified across IEEE Xplore, PubMed, Scopus, Web of Science, the ACM Digital Library and arXiv, 26 primary studies met the inclusion criteria and are analysed in depth. We organise the literature into a taxonomy covering convolutional, transformer, self-supervised, vision-language, graph-based and multimodal architectures, and we compare reported performance on the five official competition pathologies (Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion). We further examine how each study handles CheXpert's uncertainty labels, model calibration, robustness to image perturbations and explainability. The synthesis identifies persistent open problems—noisy and uncertain labels, demographic bias, weak external generalisability, the absence of standardised evaluation protocols, and limited clinical validation of reported gains—and outlines concrete future directions including foundation-model adaptation, federated training, fairness auditing, calibration-aware learning and prospective clinical evaluation.

Keywords: ChexPert, Pulmonary Disease, Image Processing, Multi-Class Classification, Medical Image Analysis, Chest X-Ray, Radiology

1. INTRODUCTION

In pulmonary medicine, x-ray images of thorax are among the most popular and easily accessible imaging modalities. Their affordability, ease of acquisition, and ability to visualize various lung pathologies make them cornerstone tools for initial evaluation and disease monitoring. However, subtle disease presentations, overlapping findings, and image variability can lead to misdiagnosis, delayed diagnosis, or inaccurate classification and detection of pulmonary disease from chest x-rays, which seems challenging, especially for less experienced radiologists[1]. For this reason, application of computer-aided diagnosis systems in hospitals is attractive possibility to increase productivity and efficiency in the interpretation of these diagnosis, through supporting with deep learning (DL). With the increasingly high efficiency and accuracy of innovative computational methods, especially deep-learning algorithms, to automatically classify and detect pulmonary diseases from chest X-ray images, in recent years, interest in this area of study has grown. This helps radiologists diagnose the specific disease accurately and promptly, and it is gradually being considered as a preliminary diagnostic tool. As mentioned in [2], DL would be highly valuable for clinicians to utilize their resources and time effectively, considering that cardiothoracic and pulmonary abnormalities are among the main causes of morbidity and mortality, according to [3]. The most used radiographic technique for lung diagnosis is chest radiography, which is also the oldest. Typically, two planes or projections are used for chest radiography. Approximately 70% of the lung capacity

is visible in at least one projection, where all anatomical structures in the X-rays' path are overlaid on one another [4]. Different types of lung abnormalities can be addressed with the help of chest x-ray as follows:

1.Consolidation: When dense materials fill the alveoli and the very small airways of the lung, it is said to be consolidated. The air passages can become blocked with material other than pus (e.g. fluid - pulmonary oedema, blood - pulmonary haemorrhage, cells - cancer) and it is important to remember that consolidation of lungs does not always indicate an infection. A region of the lung that has been consolidated turns white and thick. They are comparatively low density (blackier) if the major airways are spared. The term "air bronchogram" refers to this event, which is a typical indicator of consolidation.

2.Pneumothorax: Pneumothorax is the term for partial or complete lung collapse because of trapped air in pleural space, is known as pneumothorax. When a tension pneumothorax occurs, a chest X-ray may show a visible pleural line, no lung marking beyond line, and mediastinal shift away from affected side.

3.Pleural Effusion: It is fluid buildup in pleural area that surrounds lungs. Chest X-rays can show blunting of costophrenic angles, meniscus sign, or mediastinal shift indicative of pleural effusion.

4.Pulmonary Fibrosis: A progressive lung disease characterized by forming scar tissue in the lungs. Reticular opacities, traction bronchiectasis, and honeycombing associated with pulmonary fibrosis may be seen on chest x-ray.

5.Pneumonia: Lung tissue inflammation is typically brought on by bacteria, viral or fungal infections. Chest X-rays can reveal areas of consolidation, infiltrates, and opacities in the lungs, indicative of pneumonia.

6.Tuberculosis (TB): A bacterial infection that predominantly impacts lungs, however, may additionally impact other body parts. Chest X-ray may show findings such as nodular opacities, cavitations, and upper lobe infiltrates in active TB cases.

7.Pulmonary Edema: Fluid buildup in lungs, mostly because of fluid overload or heart failure. Diffuse bilateral opacities, Kerley B lines, and cardiomegaly suggestive of pulmonary edema can be seen on chest X-rays.

8.Pulmonary Embolism: A blockage in one of the pulmonary arteries, usually caused by a blood clot that travels from elsewhere in the body. While chest X-rays may not directly visualize pulmonary embolisms, they can show secondary symptoms including Atelectasis, Pleural Effusion, or a Hampton's hump (wedge-shaped opacity) suggestive of pulmonary infarction.

9.Interstitial Lung Disease (ILD): A set of disorders marked by inflammation and scarring of the lung interstitium. Chest X-rays may demonstrate reticular opacities, ground-glass opacities, honeycombing, or Fibrotic changes indicative of ILD.

10.Chronic Obstructive Pulmonary Disease (COPD): A group of progressive lung condition, marked by airflow obstruction and breathing difficulties, such as chronic bronchitis and emphysema. Chest X-rays can detect hyperinflation of the lungs, flattened diaphragms, and signs of pulmonary hyperinflation associated with COPD.

Recently, DL has been seen to possess tremendous potential in medical imagery for detection of diseases like lung illness. It is predicated on the composition and capabilities of the human brain. DL techniques have garnered significant attention and demonstrated remarkable potential in identifying and categorizing pulmonary conditions from several medical imaging. Application of DL in pulmonary disease detection and classification offers several advantages, including:

1.Automated Detection: DL models can automate the detection of pulmonary abnormalities from chest X-ray images, assisting radiologists in identifying suspicious findings such as nodules, infiltrates, and opacities. This automation can enhance the efficiency of radiological interpretation and expedite the diagnosis process.

2.Improved Accuracy: DL algorithms, particularly Convolutional Neural Networks (CNN), excel at learning complex patterns and features from medical images. By leveraging large-scale datasets, DL models have the potential to out-perform humans in some tasks by achieving high levels of accuracy in identifying and classifying the presence of lung illnesses.

3.Early Disease Detection: Early detection of pulmonary illnesses is made possible by DL models' capacity to spot minute abnormalities in chest X-ray imaging, which promotes timely interventions and improved patient outcomes. Early detection is crucial for conditions such as TB and lung cancer, in which early diagnosis significantly impacts treatment efficacy and prognosis.

4.Multi-class Classification: From chest X-ray pictures, DL models are capable to categorize various pulmonary conditions, e.g. lung cancer, pneumonia, TB, and COPD. These multi-class classification capabilities enhance the versatility of DL-based diagnostic systems and broaden their clinical utility across different disease entities.

5.Personalized Medicine: DL models can learn from diverse patient populations and adapt their predictions based on individual patient characteristics, including age, gender, and comorbidities. This customized approach to pulmonary disease detection and classification enables tailored treatment strategies and improves patient care outcomes.

6.Integration with Clinical Decision Support Systems: By giving real-time feedback and recommendations based on chest x-ray findings, these systems improve clinical decision-making and

promote collaborative care delivery. DL algorithms can be incorporated into clinical decision support systems to help healthcare providers make diagnostic and therapeutic decisions.

7.Scalability and Generalization: DL models might be trained using large-scale datasets that contain a range of chest X-ray images from different demographics and imaging settings. This scalability enables DL algorithms to generalize to unseen data and exhibit robust performance across healthcare settings and geographic regions.

Hence, DL is rapidly evolving into state-of-the-art (SOTA), enhancing efficacy in several medical-related applications. Consequently, these developments let physicians effectively identify and categorize certain medical conditions [5]. Cardiothoracic and pulmonary abnormalities were routinely detected and classified with Convolutional Neural Networks (CNN) because of its outstanding capacity to process data containing a high degree of spatial relationships. The literature contains numerous studies that use DL approaches to detect lung illnesses. There aren't many survey studies that examine SOTA work on this subject [6] [7] [8][9]. However, these surveys do not address different disease classifications on a single database with other techniques. In our work, we provide a comprehensive analysis of various lung diseases that can be detected through Chest X-rays with help of DL applications on the CheXpert Database. Along with discussing open difficulties, the work offers a number of potential future directions. When choosing the best DL approaches and future directions for research and development, developers and researchers can benefit from these discoveries. The scope of the systematic review will be as following:

- 1.Present a classification framework for state-of-art deep learning methods used in lung disease diagnosis.
- 2.Present an overview of current research trends in the area.
- 3.Discuss outstanding research gaps and merging opportunities for future work in this area.

The literature selection was conducted according to PRISMA 2020 statement, and search strategy. Eligibility criteria and screening process are detailed in Section 2. To summarize, 197 records were retrieved from IEEE Xplore, PubMed, Scopus, Web of Science, the ACM Digital Library and arXiv, with 26 main studies meeting the inclusion criteria, and the results of these studies were used for qualitative synthesis in Section 6.

The structure of the review is discussed henceforth. Methods for review and PRISMA flow are outlined in section 2. Section 3 provides a brief introduction to DL and SOTA architectures most relevant to medical image analysis (MIA). Section 4 introduces some of the publicly available CXR datasets and summarizes the canonical pipeline for DL using CXRs. Section 6 presents a comparative analysis of the 26 reviewed studies on the CheXpert dataset. Section 7 presents cross-cutting results and conclusions, Section 8 presents open challenges and future directions, and Section 9 gives the conclusion.

2. REVIEW METHODOLOGY

PRISMA 2020 statement and, when appropriate CLAIM checklist for studies of AI in medical imaging were followed in conducting this review.

2.1Research questions

RQ1. Which deep-learning architectures and learning paradigms have been applied to the CheXpert dataset for pulmonary-disease classification?

RQ2. How do these methods handle CheXpert's uncertainty labels and class imbalance?

RQ3. What performance has been reported on the five official CheXpert competition pathologies, and how comparable are these results across studies?

RQ4. What are the principal open challenges and future research directions?

2.2Information sources and search strategy

Searches were done in IEEE Xplore, PubMed, Scopus, Web of Science, the ACM Digital Library and arXiv for records published between January 2019 (the year CheXpert was released) and March 2026. The following Boolean query was adapted to the syntax of each database:

("CheXpert") AND ("chest x-ray" OR "chest radiograph" OR "CXR") AND ("deep learning" OR "convolutional neural network" OR "transformer" OR "self-supervised" OR "vision-language") AND ("classification" OR "detection" OR "diagnosis")

2.3Eligibility criteria

Inclusion:

- 1.Peer-reviewed journal or conference papers, or arXiv preprints with ≥ 10 citations
- 2.CheXpert used as a primary training, validation or evaluation dataset;
- 3.The task is the classification or detection of one or more thoracic pathologies;
- 4.The deep-learning methodology is described with sufficient detail for reproducibility

Exclusion:

- 1.Studies that only segment or generate images without classification;
- 2.CT-only or non-thoracic studies;
- 3.Surveys, editorials and abstracts;
- 4.Non-english publications;
- 5.Studies that mention CheXpert only in the related-work section.

2.4 Study selection and quality assessment

Records were de-duplicated and screened by title, abstract, followed by full-text screen-ing. Independent screening was done by two reviewers, and differences were settled by consulting a third reviewer. The methodological quality was evaluated by a CLAIM-based 12-item checklist covering data partitioning, label-uncertainty handling, baseline comparisons, statistical reporting, code availability and external validation.

2.5 Data extraction

For every study that was included; backbone architecture, pre-training data, uncertainty-handling strategy, loss function, data augmentation, training/valida-tion/test split, target pathologies, reported metrics (AUC, F1, accuracy), code availability and stated limitations were extracted. The complete extraction template is provided as supplementary material.

2.6 PRISMA flow

Figure 1 provides details regarding identification, screening, eligibility and inclusion steps. From 197 records identified across the six sources (IEEE 71, Scopus 48, PubMed 32, Web of Science 24, arXiv 18, ACM 4), 41 duplicates were removed. Of the 156 records screened by title and abstract, 92 were eliminated. Of the 64 full-text articles assessed, 38 were eliminated (non-primary CheXpert use 19, no classification task 9, insufficient methodological detail 7, non-English 3), leaving 26 studies included in the qualitative synthesis.

3. BRIEF INTRODUCTION OF DEEP LEARNING AND STATE-OF-THE-ART MODELS

Deep Learning is subset of machine learning that is inspired by structure and func-tioning of human brain, called artificial neural networks (ANN). Neural network (NN) algorithms need to be trained as a prerequisite to learn from and make predictions

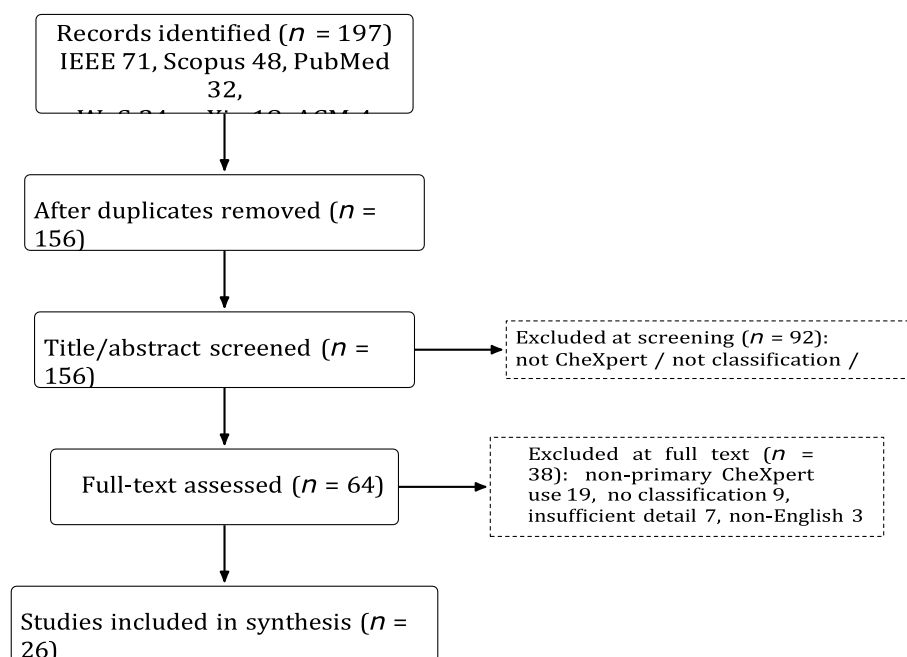


Fig. 1: PRISMA 2020 flow diagram of study identification, screening, eligibility and inclusion.

about data. As input data flows through the network, DL models' variety of layers consisting of interconnected nodes, or neurons, digest it hierarchically and extract increasingly abstract properties.

One of deep learning's key advantages is its ability to automatically recognize intricate patterns and representations in vast volumes of data without need for explicit feature engineering.

This makes it particularly powerful for tasks like medical image analysis. Deep Learning has significantly contributed to medical image processing, revolutionizing various aspects of healthcare. Some key applications include:

1. Medical Image Segmentation: Deep Learning algorithms can accurately segment anatomical structures and abnormalities in medical images like MRI, CT, and X-ray scans. Segmentation helps in precise localization and measurement of tumors, organs, or other structures, which aids in diagnosis and planning of treatment.

2. Disease Detection and Diagnosis: Deep Learning models can automatically detect and classify disease from medical images, such as detecting tumors in mammograms, identifying abnormalities in chest X-rays (e.g. pneumonia), or diagnosing diabetic retinopathy from retinal scans. These tools can aid radiologists evaluate images more effectively and precisely.

3. Image Reconstruction and Enhancement: Deep Learning techniques can advance the quality of medical images by lowering noise, improving resolution, and reconstructing missing or degraded parts of the image. This can help produce clearer and more detailed images, facilitating better diagnosis and treatment.

4. Image Registration: Medical photos from several modalities or time points can be precisely compared and analyzed via Deep Learning algorithms. Image Registration is essential for monitoring disease progression, treatment planning, and image-guided interventions.

5. Personalized Treatment Planning: Treatment outcomes and responses can be predicted by deep learning models trained on massive databases of clinical data and medical imagery. This personalized approach can aid healthcare providers in devising optimal treatment plans tailored to individual patients, which will improve results and lower medical expenses.

6. Surgical Assistance: Deep learning algorithms can provide real-time guidance and assistance to surgeons during procedures by analyzing medical images and providing augmented information, such as highlighting critical structures, tumor boundaries, or potential risks.

Medical image analysis has made use of different complex deep-learning models. Some of the notable models are as follows:

1. U-Net: U-Net is CNN architecture designed for biomedical image segmentation. Tumor segmentation in MRI and CT scans, cell segmentation in microscope pictures, and other activities have made extensive use of it.

2. DeepLab: DeepLab is a state-of-the-art CNN architecture for semantic image segmentation. It utilizes atrous convolution (also known as dilated convolution) to capture multi-scale contextual information, making it well-suited for organ segmentation in medical images.

3. V-Net: V-Net is a 3D extension of U-Net architecture, tailored explicitly for volumetric medical image segmentation tasks such as segmenting organs or tumors in 3D MRI or CT scans. Its 3D nature allows it to capture spatial dependencies more effectively.

4. 3D CNNs: 3D CNNs have gained popularity for processing volumetric medical images directly, without the need for pre-processing steps like slice extraction. Models like VoxResNet and 3D U Net have been successful in tasks like brain tumor segmentation and lung nodule detection.

5. Transformer-based Models: Transformer architectures, initially developed for natural language processing tasks, has recently been adapted for image analysis. Models like Vision Transformer (ViT) and DeiT (Data-efficient Image Transformer) have shown competitive performance in tasks such as Image Classification and object detection, and they could potentially be applied to medical image analysis as well.

6. GANs: GANs (Generative Adversarial Networks) are used for image generation and tasks like image-to-image translation and data augmentation. In medical imaging, tasks such as generating synthetic medical images for training purposes, domain adaptation, and data augmentation have all been accomplished with GANs.

7. Attention mechanisms: Similar to how humans pay attention to important details, attention systems concentrate on particular, pertinent portions of an image. In medical image analysis, where detecting minute anomalies is crucial, this is very helpful. CNN and transformer architectures are incorporating attention based models to increase the precision of tasks like disease detection and localization.

DL algorithms have been formulated to classify lung diseases and interpret chest X-rays (CXR) [?] [10] [11] [12]. These methods help radiologists, minimize diagnosis errors, identify skeptical areas, assess common abnormalities or doubts, and get around human bias and perception. A radiograph may occasionally show many pathologies, including TB, lung cancer, pulmonary nodules, edema, etc. From single radiograph or single sickness, automated DL-based system can identify several diseases.

There are many reviews on MIA that have incorporated DL models [13] [14] [15] and CXR analysis deep learning [16] [17] [18]. These reviews, though, are highly individualized and concentrate on individual illnesses such as pneumonia [19], lung cancer [20], tuberculosis [21], pulmonary nodule detection [22], and COVID-19 detection [23]. Multi-disease and single-disease chest pathology detection systems must therefore be reviewed using a thorough literature evaluation methodology framework. The evaluation and synthesis of the most advanced single-disease and multi-disease chest pathology detection systems, especially DL, is the main focus of this article.

4. THE STANDARD PIPELINE FOR DEEP-LEARNING-BASED CXR ANALYSIS

The following phases make up a deep learning-based chest pathology detection system's typical workflow:

1. Data Collection and Preprocessing

(a) Data Acquisition: A sizeable set of Chest X-ray images that have been labeled to show whether or not various chest pathologies are present.

(b) Data Cleaning and Pre-processing: Pre-process the images to ensure uniformity in size, resolution, and format. This may involve resizing, normalization, and augmentation techniques to enhance diversity and quality of dataset.

2. Model Selection or Design

(a) Model Selection: Selection of appropriate deep learning architecture for the task, considering factors

such as complexity of pathology, dataset size, and computational resources. Common choices include ResNet, DenseNet, or custom architectures tailored to specific requirements of chest pathology detection.

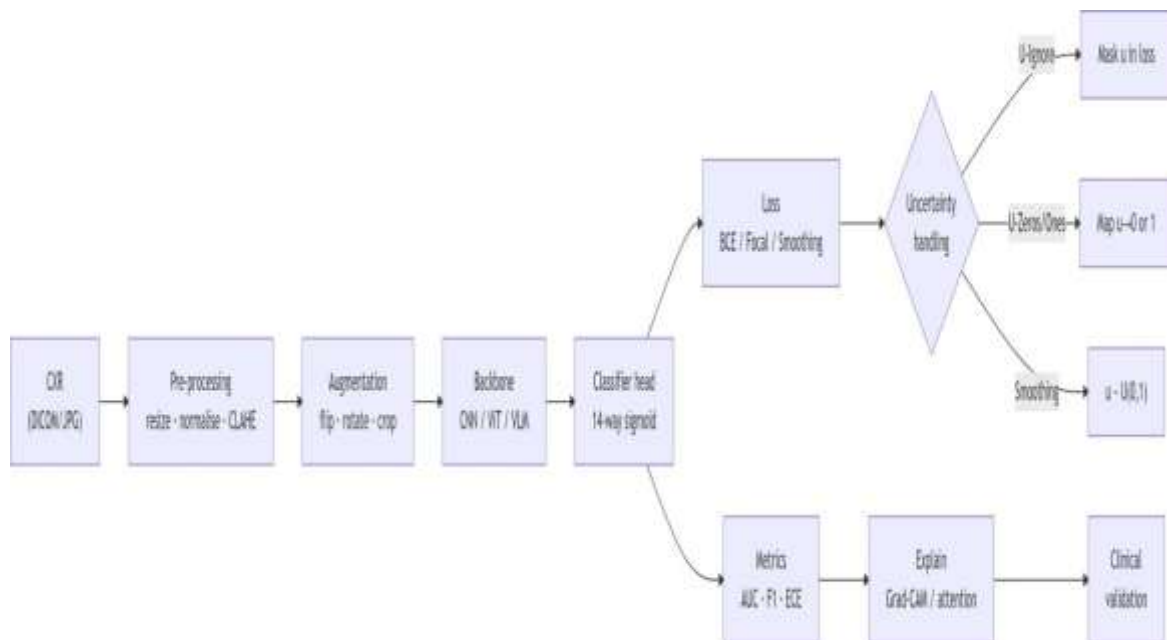


Fig. 2: End-to-End CXR DL Pipeline.

(b)Transfer Learning: The model can be initialized by pre-training its weights on large dataset to leverage learned features and accelerate training. Fine-tuning model on target dataset helps adapt it to specific characteristics of chest pathology images.

3.Training

(a)Dataset Splitting: Categorize dataset into training, validation, and test sets to evaluate model performance.

(b)Training: Train deep learning model on training set with methods like SGD, back-propagation, and optimization algorithms (e.g., Adam). To avoid over-fitting, track performance on the validation set and modify hyper-parameters as necessary.

(c)Regularization: Apply regularization techniques eg. dropout and weight decay to avoid over- fitting and improve regularization.

4.Evaluation

(a)Performance Metrics: Assess trained model's capability to recognize chest diseases on test set using performance metrics (accuracy, precision, recall, F1-score, and AUC).

(b)Clinical Validation: Validate model's perfor- mance in clinical setting by comparing its predic- tions with ground truth annotations from radiologists or other medical professionals.

5.Deployment

(a)Integration with Clinical Workflow: Integrate the trained model into clinical workflow systems to assist radiologists in chest pathology detection during routine image interpretation.

(b)User Interface: Develop user-friendly interface for accessing model's predic-tions and integrating them into radiology reports or electronic health records (EHRs).

(c)Continued Monitoring and Maintenance: To guarantee accuracy and reliability, track the model's performance in real-world scenarios and make necessary updates.To guarantee accuracy and reliability, track model's performance in real-world scenarios and make necessary updates.

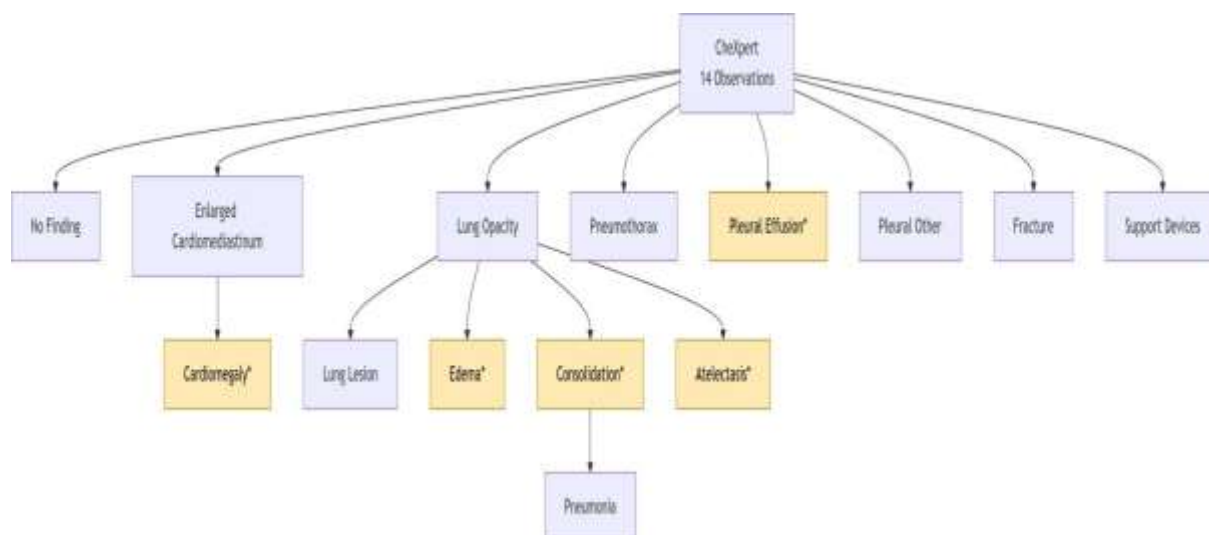


Fig. 3: Dataset Hierarchy.

5. DATASETS AND APPLICATION OF DEEP LEARNING TO CXR

5.1 Comparison of publicly available CXR datasets with multi-label, multi-disease annotations

Several large-scale, publicly available chest-radiograph collections underpin contemporary deep-learning research. Table 1 summarises the principal datasets that overlap with the CheXpert label space and are most often used either as auxiliary training corpora or as external validation sets in the studies reviewed in Section 6.

CheXpert [24] is distinguished by its rule-based labeller, which extracts 14 observation labels from free-text radiology reports together with explicit uncertainty markers; this property motivates much of the methodological work surveyed here. MIMIC-CXR-JPG [25] shares the same label schema, enabling cross-dataset transfer studies. PadChest [26] contributes Spanish-language reports and a richer hierarchical label set, while ChestX-ray14 [27] provides bounding-box annotations for a small subset, supporting weakly supervised localisation. OpenI [28] and VinDr-CXR [29] are smaller but high-quality, expert-annotated test beds that are increasingly used for the external validation of CheXpert-trained models. The PLCO collection [30] adds a longitudinal screening perspective.

5.2 Overview of Deep Learning

Hubel and Wiesel [31] studied cat visual cortex cells and discovered that the animal's visual nerve system identifies objects in layers, which led them to propose the idea of receptive field in 1962. A basic visual processing framework was developed. When an object passes through a visual area, the brain uses edges to create highly advanced visual information. This model was not intended to be actual model, even though it aids in understanding how brain functions. Based on idea of receptive field, Japanese researcher Fukushima presented a novel cognitive machine (neocognitron) in 1980. It may be considered as first implementation network of CNN and first application of idea of receptive field in ANN [32]. Rumelhart et al. proposed famous back-propagation algorithm, which is capable to train NN with two hidden layers and transmit errors in reverse for changing weight of network [33]. Hecht-Nielsen suggested that 3-layer network of hidden layers may approximate a continuous function in any closed interval, suggesting that multilayer perceptrons have the capacity for "universal approximation" [34]. LeCun proposed that LeNet solves the issue of handwritten

Table 1: Comparison of publicly available chest-radiograph datasets used in CheXpert-related studies.

Dataset	Images	Patients	Labels	Views	Year	Salient features/licence
CheXpert [24]	224,316	65,240	14	F/L	2019	Rule-based labels with explicit uncertainty; Stanford research licence.

MIMIC-CXR-JPG [25]	377,110	65,379	14	F/L	2019	Paired free-text reports; PhysioNet credentialed access.
PadChest [26]	160,868	67,625	174	F/L	2020	Hierarchical Spanish labels; 27% expert annotated.
ChestX-ray14 [27]	112,120	30,805	14	F	2017	880 bounding-box annotations; CC0 licence.
OpenI [28]	7,470	3,955	MeSH	F/L	2016	Manual MeSH/RadLex coding; public domain.
VinDr-CXR [29]	18,000	18,000	28	F	2022	Triple expert annotation with bounding boxes; PhysioNet.
PLCO [30]	>185,000	154,901	22	F	2013	Cancer-screening trial cohort; controlled access.

F = frontal (PA/AP), L = lateral.

numeral recognition. Convolution, pooling, and full connection layers have now been identified as the fundamental components of CNN's design [35].

Deep Belief Network suggested in 2006 that neural network training becomes progressively easier [36]. For layer-by-layer training, a greedy approach was employed, which effectively improved performance. NN model's architectural size has expanded along with the amount of data. AlexNet may be regarded as landmark algorithm [37]. Any size image can be used as input for processing once neural network's processing capacity has progressively increased. VGG-Net [38] GoogLeNet [39] was proposed in 2014, and ResNet [40] deepened more layers than previous network of AlexNet. Dense neural networks [41].

Because transfer learning makes it possible to create precise models, it has been extensively employed approach in computer vision [42]. A model that has been learned in one domain can be applied to another using transfer learning. It is possible to do transfer learning with or without pre-trained model. To begin with, a model is created for another activity. Even if pre-trained model is trained on separate task, learnt features are typically demonstrated to be helpful for current task.

In training DL model, the goal is to get right weights to network via number of forward and backward iterations. Models trained on big datasets can be pre-trained and deployed to current problems using weights and architecture they have learned. The lower cost of training new model is one benefit of applying pre-trained model [43]. This is because it only needed to learn the weights of the final few levels because it was making use of pre-trained weights.

Many popular CNN models are initially trained on ImageNet before being adapted to specific applications [44]. The ImageNet collection was assembled from images obtained online, with labels assigned by human annotators through Amazon's Mechanical Turk crowdsourcing service. The transfer learning methods are categorized into two groups: fine-tuning and CNN as a feature extractor. Fine-tuning a pre-trained CNN typically involves retaining the learned knowledge in some layers and updating others so that the model can perform better in new task. [45]. It's a common practice in Transfer Learning to preserve the parameters of the first few layers and retrain the final layers. Since, the initial layers learn common visual features like edges, textures and colour regions. These learned features are capable enough to be transferred to other image classification tasks. During fine-tuning, the high-level features learned by pre-trained models are further optimized to extract high-level patterns and structures from new dataset. Finally, CNN is used as a feature extractor, on the other hand. This is achieved by removing the last fully connected layer (that computes the probabilities of the 1000 classes in ImageNet) and the trained model can be used as a fixed feature extractor, enabling the extraction of relevant features from the new dataset.

In ensemble learning techniques, predictive performance improves by combining the predictions of several classifiers rather than relying on a single model [46]. Ensemble reduces the fluctuations in predictions, thus increasing the accuracy of predictions beyond that of any single model. Ensemble learning can be implemented through techniques such as majority-vote aggregation, averaging prediction probabilities and stacking, where the outputs of several classifiers are combined to obtain a final prediction. During validation, every classifier predicts a class for each sample. The final prediction corresponds to the class label that receives the majority number of votes from the ensemble members. There are some models whose votes are more important than others, therefore there's an alternate form of majority voting, called weighted majority. Under the probability averaging strategy, each model contributes its prediction score, which are averaged and the class with the highest mean score is selected as the final prediction [47]. One of these is the weighted averaging that weights the score given by each model and calculates the average. The stacking ensemble algorithm utilizes predictions of base models as inputs to a higher-level model which learns the most effective way to combine them and improves overall predictive performance [46]. Self-supervised learning (SSL) is recently developed method in field of AI, a specialty of MIA. It offers a promising way to make use of huge amounts of unlabeled medical imaging data to enhance diagnosis, prognosis, and treatment planning. SSL uses labels that are not annotated by humans to train a model to solve "pretext" task. The fundamental aim of this training is to acquire some attributes and representations that are useful for "downstream" task, which in this case will be the actual task being solved by the training. In MIA, pretext task could be predicting part of an image given the rest (e.g., inpainting) or predicting rotation applied to an image. The downstream task could be a clinical task such as disease detection or segmentation of a particular organ. Because of the high cost and time needed to acquire and annotate data, supervised learning may not be a viable method in medical imaging applications; here, SSL can provide solution to limitations of supervised learning. Moreover, SSL can extract more complex and high-level features than conventional learning techniques, which could benefit performance of learned models in MIA [48][49].

5.3 Application of Disease Detection & Classification using Deep learning

The process of detecting lung diseases using deep learning generally involves the following steps:

1.Data Collection: The first step involves gathering a large amount of lung imaging data. This could be Chest X-rays or CT scans, depending on the diseases to be detected.

2.Data Pre-processing: This gathered data should be pre-processed to make sure that it's in an appropriate structure for the deep learning model. Image contrast can be improved using CLAHE, making important structures and details more distinguishable [50]. To improve diagnostic accuracy, image enhancement procedures such as extracting lung regions and suppressing bone structures can be implemented to localize relevant anatomical areas. Once the region of interest is identified, disease detection can be carried out more effectively. Moreover, edge detection algorithms can be employed to obtain alternative image representations that support feature extraction and classification [51][52][53]. The images can be supplemented with data to enrich available data. Another way of feature extraction is to provide the deep learning model with specific features that it can detect and classify a specific object or class. This stage produces a series of images with either better image quality or the removal of undesired elements. This step's output is the improved/modified images to be used for training purposes. This could be scaling images, normalizing pixels, adding variations to the dataset etc.

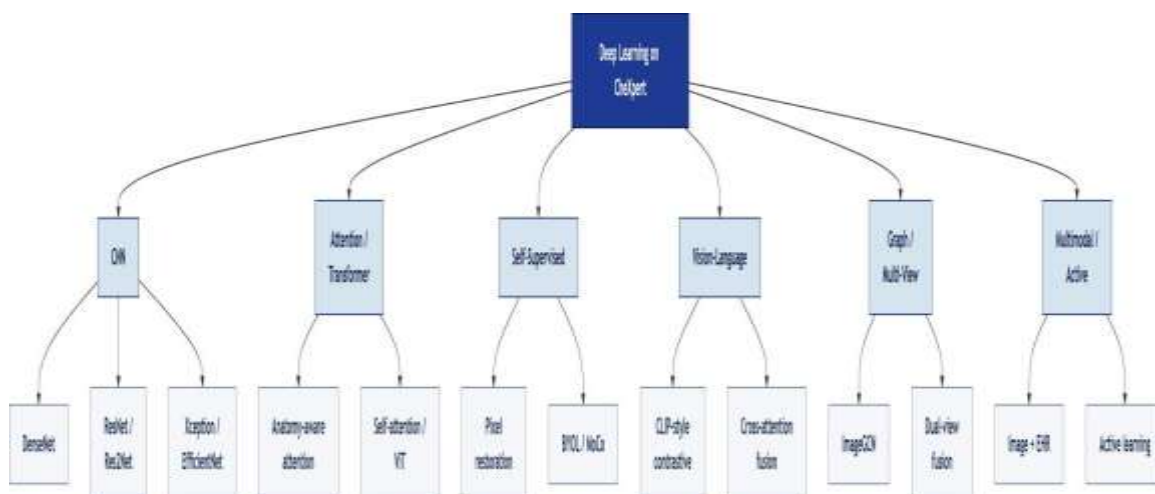


Fig. 4: Taxonomy of deep-learning approaches applied to the CheXpert dataset, derived from the 26 included primary studies.

3.Labeling: If supervised learning is used, each image in the dataset must be labeled with the correct diagnosis. Medical professionals typically do this, and it can be a time- consuming process.

4.Model Selection: A suitable deep-learning model needs to be selected. Convo-lutional Neural Networks (CNNs) are often used for image analysis tasks, but other architectures like ResNet, DenseNet, or U-Net might be used.

5.Training: The pre-processed dataset, together with its corresponding labels, is then used for model training and parameter optimization. In the course of training, the model gains the ability to recognize features in images indicative of the diseases.

6.Validation and Testing: The trained model is subsequently evaluated using independent validation and testing datasets to assess its performance and gener-alization capability. This evaluation is performed to assess the model's ability to perform reliably on new and unseen datasets.

7.Evaluation: The effectiveness of model is measured through performance indi-cators, such as accuracy, precision, recall, or the Area Under the ROC Curve (AUCROC).

8.Deployment: If the model performs well, it can be deployed in a clinical setting. Here, new lung images can be analyzed to detect the presence of disease.

9.Monitoring and Updating: The model's performance should be continually monitored once deployed. If the new data becomes available, the model can be retrained or fine-tuned to improve its performance.

6. ANALYSIS AND COMPARISON OF RECENT TRENDS IN LUNG- DISEASE CLASSIFICATION ON CHEXPRT

This section synthesises the 26 primary studies retained after PRISMA screening (see Section 2). Figure 4 presents the taxonomy of deep-learning approaches used to organ-ise the corpus, and Figure 5 shows the distribution of publications by year. Table 3 provides a head-to-head summary of architecture, learning paradigm, uncertainty han-dling, reported performance and key contribution for every included study; Table 4 aggregates the uncertainty-handling strategies adopted across the corpus, and Table 2 compiles per-pathology AUC values for the studies that report them. The remainder of the section discusses each study in more detail, grouped chronologically.

Zhou, Hong-Yu., Lu, Chixiang. et al. [61] suggest a novel unified SSL framework, PCRLv2, to overcome shortcomings of comparative SSL in MIA. Auhors introduce multitask optimization approach on the feature pyramid, consisting of pixel restoration

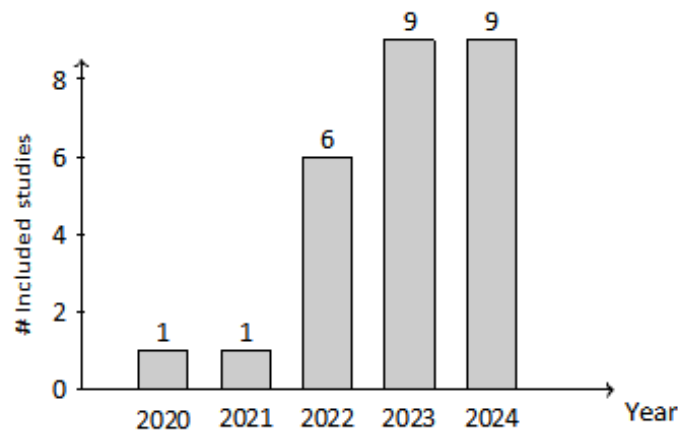


Fig. 5: Distribution of the 26 included primary studies by publication year (2020–2024). The acceleration after 2022 coincides with the rise of self-supervised and vision–language approaches.

Table 2: Per-pathology AUC reported for the five official CheXpert competition labels. Values are taken verbatim from the source papers; “–” indicates the value is not reported. Atel. = Atelectasis; Card. = Cardiomegaly; Cons. = Consolidation; Edema = Edema; PIEff. = Pleural Effusion.

Study	Atel.	Card.	Cons.	Edema	PIEff.	Mean (5)
Irvin et al. (CheXpert baseline) [24]	0.858	0.832	0.899	0.941	0.936	0.893
Pham et al. (Conditional) [54]	0.825	0.855	0.937	0.930	0.923	0.894
Kamal et al. (Anatomy-XNet) [55]	–	–	–	–	–	0.870
Termritthikun et al. [56]	–	–	–	–	–	0.871
Chong et al. (best ensemble) [57]	–	–	–	–	–	0.882
Zhang et al. (LCFL, 14 labels) [58]	–	–	–	–	–	0.765†
Rocha et al. (STERN, normal/abnormal) [59]	–	–	–	–	–	0.856‡
Tiu et al. (CheXzero, zero-shot) [60]	0.811	0.901	0.897	0.907	0.929	0.889

and siamese feature comparison at different scales. Key points of paper's framework include:

1.Self-supervised learning:The approach utilizes SSL, where a model learns from inherent structure of data without explicit human annotations. This method is particularly beneficial in medical imaging applications, where generating high quality labeled datasets can be both challenging and expensive due to the need for expert annotation.

2.Visual Information Preservation: The framework emphasizes preserving visual information during pre-training. Designing tasks that encourage the model

Table 3: Summary of deep-learning studies evaluated on the CheXpert dataset (2020–2024). AUC5 denotes the mean AUC over the five official competition pathologies (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion); NR = not reported in the source paper.

Study	Backbone	Paradigm	Uncertainty handling	AUC5	Key contribution / limitation
Zhou et al. [61]	Non-skip U-Net	Self-supervised (PCRLv2)	U-Ignore	NR	Pixel restoration + Siamese; limited to feature-pyramid pre-training.
Costa et al. [62]	VGG-16	Supervised CNN	U-Zeros	NR	Compliance / view classification rather than disease per se.
Meshnick et al. [63]	Xception	Multimodal CNN	U-Ones	NR	Image + binary metadata fusion; small-scale validation.
Termritthikun et al. [56]	EEEE-Net-C2 (DenseNet-161 teacher)	Knowledge distillation	U-Ones	0.871	On-device deployment with XAI; 5-pathology AUC only.
Nie et al. [64]	Transformer + SCM	Causal / IV learning	U-Ignore	NR	EHR fusion via instrumental variables; complex training.
Liu et al. [65]	DenseNet-121	Curriculum (ML-LGL)	U-Ones	NR	Local-to-global ordering; sensitivity to abnormality ordering.
Li et al. [66]	Res2Net + ECA	Multi-scale CNN	Biased Focal Loss	NR	Multi-level residual fusion; no external validation.
Chong et al. [57]	28 CNN/EffNet variants	Robustness study	Label smoothing	–	Perturbation robustness benchmark across 28 backbones.
Talasila et al. [67]	ResNet-50 (BYOL)	SSL + weak supervision	U-Ignore	NR	Grad-CAM auxiliary loss for interpretability; small NIH pre-training.
Rabbani & Bartoli [68]	Stanford Baseline	Noisy-label learning	UCA confidence head	NR	Joint noisy-label + selective prediction; no external test.
Kamal et al. [55]	DenseNet-121	Anatomy-aware attention	U-Ignore	NR	Anatomy-prior attention; depends on segmentation quality.
Ni et al. [69]	Dual-stream CNN	Multi-view fusion (VAF / 3DRF)	U-Ones	NR	Frontal+lateral fusion; requires paired views.
Mao et al. [70]	DenseNet + GCN	Graph-based (ImageGCN)	U-Ignore	NR	Inter-image relations; PPS reduces parameters.
Biesner et al. [71]	DenseNet + RNN	Patient-history sequence	U-Zeros	NR	Longitudinal modelling; short sequences only.

Wan g et al. [72]	LSR generator	Data- augmentat ion GAN	U-Ones	NR	Healthy-CXR synthesis for abnormality augmentation.
Chen et al. [73]	DenseNet	Multi-agent classifier	U-Ignore	NR	“Triage” selectors mitigate imbalance; marginal gain.
Crespi et al. [74]	β -VAE	Generativ e + ML head	U-Zeros	NR	Latent-feature classification; frontal views only.
Mitra et al. [75]	Xception, ResNet-18	NAS- style sweep	U-Zeros	NR	Architecture comparison; class-imbalance not addressed.
Ren et al. [76]	DenseNet-121 + EHR	Multimodal	U-Ones	NR	Age-stratified pneumonia detection in elderly.
Zhang et al. [58]	ResNet + self-attention	Label- correlation	U-Ignore	0.765	LEFL+MLCL; F1 = 0.446; competitive on 14 labels.
Mahapatra et al. [77]	VAE + GAT	Active learning	U-Ignore	NR	Multi-label informativeness score.
Shu et al. [78]	CLIP-style	Vision- language	U-Ignore	NR	Multi-level contrastive; needs paired reports.
Zheng et al. [79]	CLIP	pre-training	U-Ignore	NR	Strong zero-shot; lightweight prompt module.
Park et al. [80]	DINO + CXR- BERT	Weakly supervised prompt	U-Ignore	NR	Zero-shot error detection in radiology.
Rocha et al. [59]	CNN + Spatial Transformer	Cross- attention VLM Attention CNN	U-Ones	0.856	Binary normal/abnormal; reduces FP via STN cropping.

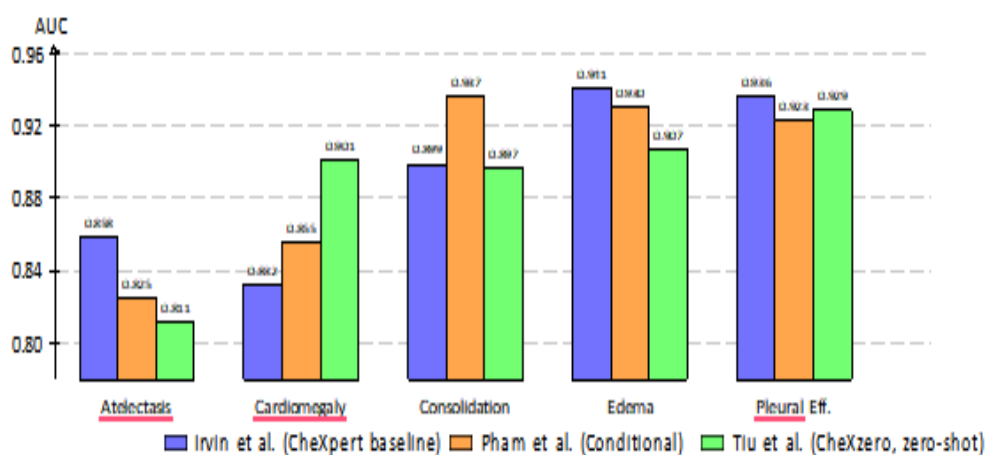


Fig. 6: Per-pathology AUC reported on the five official CheXpert competition labels by the studies that publish per-class values (Table 2). Edema and Pleural Effusion are consistently the easiest classes; Atelectasis and Cardiomegaly remain the hardest. CheXzero matches the supervised baseline on Cardiomegaly without any CheXpert labels at training time.

Table 4: CheXpert uncertainty-label handling strategies adopted by the reviewed studies.

Strategy	Description	# Studies
U-Ignore	Treat u labels as missing during loss computation.	12
U-Zeros	Map $u \rightarrow 0$ (negative).	5
U-Ones	Map $u \rightarrow 1$ (positive).	6
Label smoothing	Replace u with $U(0,1)$ or fixed smoothed value.	2
Biased focal loss	Re-weight uncertain / hard negatives.	1

to retain relevant features across different views or modalities of the same image aims to improve the model's generalization capabilities.

Recent advances in SSL for Computer Vision is based on contrastive learning techniques. In this method models learn meaningful and robust feature representations by comparing different views of same image using siamese networks. But preserved high-level semantics lack sufficient local information which is important for MIA. To solve locality issue of comparative SSL, the authors proposed incorporating task of pixel restoration to encode more pixel-level information explicitly into high-level semantics. Authors discussed preserving scale information, which is powerful tool in understanding images as well. The resulting framework is posed as multitask optimization problem over a feature pyramid. In order to construct feature pyramid and create a sub-crop to take role of multi-crop in 3D medical imaging, a different design was put out as a non-skip U-Net M. Costa, S. C. Pereira, et al. [62], in their paper, suggested a method to identify non-compliant photos using deep features taken from patient position classification model and a pre-trained VGG-16 model-based method applied to CheXpert.

David C. Meshnick et al. [63] proposed machine learning and multimodal learning techniques in analyzing CheXpert chest X-rays in their paper. Various pre-trained and non-pre-trained CNNs (ResNet50, ResNet101, DenseNet121, InceptionV3 and Xception) were evaluated and tested with various combinations of image and meta-data inputs. The results showed that inclusion of a multimodal data, such as simple binary meta-data, had a significant impact on classification accuracy over conventional unimodal methods or advanced MADDi models. This demonstrates how multimodal learning can improve image classification performance and enhance data representation. Xception models performed very well in CheXpert tests. These findings represent an improvement in performance and highlight the power of learning through multiple modalities in data-rich applied science research.

In the paper by C. Termritthikun, A. Umer et al. [56] KD framework to develop a lightweight DL model capable of performing real-time multi-label classification of chest radiographs. They evaluated different CNN and transformer architectures as teacher models and transferred their knowledge to a smaller student network. To enhance, transparency, explainable AI interpretations of model's predictions, that has been enhanced by the KD. Notably, the EEE-Net-C2 model, trained using DenseNet 161 as the teacher, achieved an AUC score of 87.1% on chexpert dataset.

W. Nie, C. Zhang, et al. [64] discussed the problems with confounding factors (e.g., image resolution or noise problems) which affected the performances of the models. To overcome these issues, authors came up with an Interpretable IV learning framework that removes spurious association and provides accurate representation. The framework begins with the construction of a structural causal model to identify confounders and learns initial representations of instrumental variables. These representations are subsequently enriched with information from Electronic Health Records and integrated using transformer based semantic module for improved feature learning.

Liu, Zean., Cheng, Yuanzhi, et al. [65] have used Curriculum Learning (CL) paradigm to propose method for abnormality classification on chest X-ray images, called ML-LGL. This is done by training successively more abnormal (local) to more ones (global) in Deep Neural Network models. The learning paradigm is based upon the gradual memorization of humans from less categories to more categories. Empirically, it has been found that Local to Global Learning can escape from local minima and improve the classification performance in tasks when simulated. LGL, however, is designed to handle multi-class classification problems where the prediction is one hot vector per image. A multi-label setting is required as more than one abnormality can be observed on chest X-ray. Therefore, LGL is extended to multi-label case and the authors suggested ML-LGL. This paper proposes 3 types of ASF: Correlation, Similarity and Frequency Based, for prioritizing abnormality and proposes learning orders similar to radiologist's progression.

Li, Qiang., Lai, Yu., et al. [66] discussed the limitations of existing methods to derive multi-scale features and precisely capture spatial characteristics of lesions while effectively addressing thoracic disorders that often occur concurrently and exhibit wide variability in lesion size. To resolve this issue, authors have suggested MLRFNetto classify thoracic diseases. By incorporating receptive fields of varying sizes, the framework effectively enriches disease relevant spatial representations and reduces the influence of regions that are not relevant to diagnosis. MLRFNet consists of two parts: A multi-scale feature extractor for learning semantic information at various scales from chest radiograph images and MRFC for encoding the

pathological information of the diseases under consideration.

The Res2Net [81] network extract information from receptive fields of different sizes to better capture lesions with varying dimensions, there by improving the model's capacity for multi-scale feature representation and analysis. To overcome the challenges of lesion size variability, MRFC framework is incorporated to enhance lesion localization and improve discriminative ability of the model. ECA enables the model to dynamically adjust the importance of different feature channels, which helps to focus more effectively on meaningful patterns within the feature maps and increase the transmission of useful feature channels including the important pathological feature channels, while suppressing irrelevant channels. To address the imbalance between positive and negative samples across disease categories, BFL is proposed to enhance emphasis on difficult negative examples during training and guide optimization process towards these difficult instances. This helps model to achieve better discriminative capability and overall classification accuracy.

C. F. Chong, X. Fang, et al. [57] examined how image distortions not encountered during training, which includes brightness changes, contrast alterations, noise, image compression and spatial transformations affect the performance of medical image interpretation models. A total of 28 Imagenet pre-trained architectures, including ResNet, DenseNet and EfficientNetV2, were fine tuned on ChexPert dataset and evaluated. To measure robustness, six types of synthetic image perturbations-brightness adjustments contrast variations, blur, noise addition, JPEG compression and spatial transformations were applied to chest radiographs across ten severity levels. The results were used to investigate the relationship between model robustness and architectural characteristics, highlighting the challenges posed by image quality variations in clinical settings. The findings of the experiment were as follows:

1. Models that performed well on ImageNet are generally more resilient to image distortions after being fine-tuned for chest radiograph classification upon chexpert dataset.
2. Achieving better chest x-ray classification accuracy after fine-tuning does not automatically guarantee stronger robustness to image distortions. Models with higher accuracy generally handle slight perturbations more effectively, but under severe image degradations, model with lower baseline accuracy may sometimes prove more resilient and maintain their performance.
3. The architecture design is having important contribution towards determining model robustness. Although ResNet models frequently deliver higher accuracy, they are often more susceptible to performance degradation under perturbations. In comparison, efficientNet and EfficientNetV2 architectures, even though their overall performance may vary considerably.

Five models were trained for each of the 28 CNN models, and then an average ensemble was built up for CXR classification. To adopt the models to the target task, the models were further optimized on chexpert dataset for ten epochs. During the training the complete set of model parameters remains unfrozen. Label smoothing regularization is used to deal with the uncertainty of the labels in the training set. That is, replacing the ambiguous names with arbitrary scalars, picked at random from a uniform distribution. To evaluate the prediction accuracy, binary-cross entropy loss function is calculated between the model's output and ground truth values. The study focused on five pathological conditions that were identified according to their occurrence rates and significance in patient diagnosis and management. These pathologies were evaluated against the performance metric mean area under the receiver operating characteristics curve.

A. Talasila, A. Karthikeyan, S. Alle, et al. [67] have talked about the non-performance of SOTA machine learning method with regard to explainability. Class activation mapping results are seldom assessed for their conciseness and anatomical correctness. Although the model performs well on their training data, they may not remain sufficiently accurate or consistent when applied to other datasets within the same modality. They are not used due to lack of an ability to explain the performance it predicts. To solve these problems, authors introduced a hybrid pipeline where the model is first pre-trained using self-supervised and weakly supervised methods and then fine-tuned with labeled data. This methodology improves robustness across datasets and incorporates an auxiliary loss to strengthen feature learning. The authors were inspired by BYOL, the framework achieves effective image representation learning through self-supervised training, removing the necessity for negative pair sampling commonly used in contrastive learning methods. To achieve robust feature extraction, the architecture employs a hybrid pre-training pipeline that integrates self-supervised and weakly supervised learning with auxiliary loss optimization. Then it is refined through supervised fine-tuning to generate more discriminative and transferable feature representations. In the pre-training stage, the model was trained on NIH CXR dataset in self-supervised manner with auxiliary loss that is derived from probability distributions of Grad-CAM heat map and the bounding boxes used to annotate the images. Finally, the supervised fine-tuning is done on CheXpert dataset and the model predicts multi-label classification probabilities. Integrating limited annotated samples into the pre-training phase of a largely unlabeled dataset contributes to improve localization capabilities and increased prediction confidence, while preserving cross dataset performance. The proposed architecture incorporates a ResNet-50 backbone encoder coupled with Grad-CAM for interpretability. Empirical findings reveal that the model produces finer and more reliable localization maps, achieves improved predictive performance and exhibits

superior transferability across heterogeneous datasets. By leveraging annotations during the pre-training stage, the framework learns where to focus within the images, resulting in richer feature representations and more effective downstream learning. This is an improvement over existing SOTA methods which apply blending at the image level.

N. Rabbani and A. Bartoli [68] presented a problem with noisy labels in the dataset and its negative impact on neural network training. Current approaches deal with either with noisy labels for training models, or with selective prediction in isolation, but do not necessarily take into account the intrinsic relationship between these two tasks. The authors relate these challenges and introduce novel approach Unsuper-vised Confidence Approximation (UCA) aiming to tackle them jointly. UCA enables training of neural networks for primary task (e.g., image segmentation and classification) with the prediction of confidence levels while training concurrently. All of this is accomplished while taking into account datasets with noisy labels. UCA incorporates a confidence estimation head into primary deep neural network, enabling the model to estimate the reliability of the primary task predictions. It was proposed by author two varieties, Global UCA and Pixel UCA head. The proposed framework employs a global average pooling layer (K) followed by fully connected hidden layers equipped with ReLU activations to enhance feature learning and representation capabilities. A sigmoid activation function was used along with a batch normalization layer in the output. This will help to estimate a single confidence per sample. The Pixel UCA is inspired by Global UCA but is lacking Global Averaging layer and has convolutional hidden layers rather than fully connected layers. Pixel-wise selective prediction is possible in the Pixel UCA. The UCA loss is designed to support the learning process of the main network and UCA head at the same time. It doesn't require confidence labels and is thus unsupervised. The baseline network proposed a model named Stanford Baseline (SB) was used as the main network, and it is extended by adding the proposed UCA extension. The SB was trained to be several binary classifiers, one for each of the diagnoses. The accuracy and AUC of the network were calculated by the average of all the binary classifiers and were impressive.

The anatomy-Xnet, an anatomy aware attention-driven network for thoracic disease classification by U. Kamal, M. Zunaed, N. B. Nizam et. al. [55]. The model uses anatomical attention mechanisms to focus on relevant spatial regions within chest images. The study utilizes semi-supervised learning to transfer knowledge from a small annotated dataset of organs to larger dataset where such annotations are unavailable. The proposed Anatomy-Xnet framework integrates a pre-trained DenseNet-121 backbone with anatomy-aware attention and probabilistic weighted average pooling module, enabling enhanced anatomical attention learning and improved thoracic disease classification. This is an Anatomy-XNet that uses both information from anatomy segmentation and CXRs formats to classify thoracic diseases. A major challenge in integrating the information in the context is that it is hard to find the right annotations for the anatomy region in contours for large-scale datasets. The training of segmentation network was followed by semi-supervised learning approach and the segmentation masks of the anatomy were acquired. The proposed architecture incorporates an anatomy-aware attention mechanism on the top of the DenseNet-121 feature extraction backbone. This helps the network to focus on anatomically significant regions. The Anatomy Aware Attention module not only helps to emphasize the importance of the anatomical location associated with thoracic disease, but also keeps information beyond anatomical masks via residual attention vector (attention vector outside the masks), which means it is less influenced by misaligned anatomical masks. Besides, the authors introduce a novel pooling operation layer—PWAP that explicitly utilizes probability attention map computed from feature activation map to boost salient regions of feature space.

In order to improve categorization, J. Ni et al. [69] established innovative dual-view DL framework that makes use of intermediate bi-directional fusion architecture to take use of intrinsic spatial correlation between two views, namely frontal view and lateral view. Authors have proposed two schemes for fusion, which are based on spatial correlation of the paired views, VAF and the 3D reconstruction fusion module which is based on spatial information is used to back project the 2D to 3D image fusion. Both fusion methods are employed between two parallel input CNN streams and are utilized to produce a signal which is used to recalibrate each input stream. To understand the impact of various configurations of suggested techniques, authors have taken into account following variations of two suggested fusion modules.

1. Networks that uses weights learned from Image Net dataset are identified as VAF-Net-Imagenet and 3DRF-Net-Imagenet.

2. The 3DRF-NetC2L-state architecture extends 3DRF-Net-C2L by removing the propagation of the accumulated signals from previous layers when calculating the activations of the current layer.

It was observed that the misaligned frontal and lateral views along the Y-axis have more chances of making an incorrect classification from the classification result obtained by VAF-NetC2L. To enhance the classification findings, the authors have suggested a dual-stream multi-image fusion framework and two fusion modules that take advantage of the in-intrinsic spatial correlation of the two correlated chest X-ray views. In order to provide recalibrating signals in two input streams so that models concentrate on crucial features to improve classification accuracy, the two fusion techniques can efficiently extract and fuse key features. The fusion approaches' outcomes are noticeably superior than those of early and late fusion

models, suggesting that spatial correlation of intermediate feature layers is important and should be considered.

C. Mao, L. Yao, and Y. Luo [70] developed ImageGCN a graph based deep-learning framework that captures interimage dependencies for chest radiograph analysis. By incorporating these relationships, the model produces more informative representations, enabling better accuracy in detecting and localizing diseases than prior approaches. The suggested ImageGCN model incorporates PPS (Partial Parameter Sharing) scheme to reduce complexity and improve disease identification and localization, results show that ImageGCN outperforms the baseline.

To advance chest x-ray analysis D. Biesner, H. Schneider et. al. [71] employed an RNN-based framework that integrates patient historical data with image analysis, leading to more effective chest x-ray classification. All experiments are based on base architecture DenseNet, pre-trained on ImageNet with a linear classifier with 5 class disease classification. It has been noticed that fewer parameters lead to less overfitting and the length of the image sequences is not sufficient enough to require the special long-term memory of LSTMs. RNN is trained using label information to use the information provided by the radiologists when looking at previous images. Each encoded image is integrated with a binary vector indicating its corresponding labels before being processed by the RNN. Training is performed using binary cross-entropy loss over 30 epochs, with a learning rate of 0.001, Adam optimizer, and reduce-plateau scheduling strategy. Poisson bootstrapping was used to produce 10 statistically valid estimates of model performance using validation and test splits. The evaluation of the metric was done on the aggregated bootstrapped datasets.

Z. Wang, X. Zhang et al. [72] discussed data augmentation technique to enhance efficacy of disease classification on chest X-ray images with abnormality augmentation. The suggested approach is to put input chest X-ray image into LSR generator, which can generate the paired healthy chest x-ray image. The input and output chest x-ray images are processed through an abnormality augmentation module, which identifies probable disease regions and strengthens abnormal signals. A computer program used for thoracic disease classification model uses the augmented chest x-ray image and obtains better performance. To solve issue of imbalanced data distribution in medical images and improve balance and robustness of classification, W. Chen, J. Feng et al. [73] proposed novel concept "TriageNet", which is end-to-end multi-agent classification network. TriageNet is composed of the diagnostic agents and triage selectors that have been optimized as an end to end training process. The "Diagnosis Agents" are, in essence, a miniature computer network that includes a selection of features from the backbone, a number of classification jobs and number of data sets for each classification job that are fed into the training process. The inner structure of "Diagnosis Agents" is a layer of fully-connected layers using the selected elements of features. After a softmax operation on the agent's output, loss function is cross-entropy for each agent. "Triage" selectors have fully connected layer of inner structure and ground truth can be synthesized based on the output of "Diagnosis Agent". The proposed architecture tested on U-Ignore results on CheXpert since the data imbalance is the highest, and the number of cases is the smallest. The DenseNet network was used as the backbone for CheX-pert. As shown in the result of the experiment, TriageNet is slightly better than DenseNet baseline.

L. Crespi, D. Loiacono et.al. [74] successfully classified chest radiograph images from latent features using β Variational Autoencoder (VAE). Feature representations obtained from chest x-ray images through deep learning approaches were passed to machine learning classifiers, leading to improved and reliable classification performance. These highly abstracted features are extracted using VAE, and ML models are now being trained for disease classification. For the experiment only frontal chest x-ray images are being considered.

A. Mitra, A. Chakravarty et.al. [75] investigated various deep CNN architectures through a structured search to improve chest radiograph screening. Various CNN architectures were tested for the classifying chest X-ray images, with Xception and ResNet-18 showing best performance. There were restrictions in Sampling size and class distribution. U-Zeros (i.e. replacing the uncertain label with no diseases) is explored for handling uncertain labels during training.

H. Ren, F. Jing et. al. [76] talked about that and discovered that the model that has not been age stratified was not very good for elder people. They introduced their model called CheXMed and showed its performance to be better than the baseline models on all the evaluation measures for detecting pneumonia in elder people. The study found that existing pneumonia detection models, such as CheXNet, DenseNet121, and ResNet18, performed unequally for old and young people, with a significant decrease in performance for the elderly. Another interesting result of the study is that the addition of medical records raised the accuracy of pneumonia detection up to 6.6% more accurately and 20.5% more recall. Proposed CheXMed is better than CheXNet in accuracy (+10.1%), precision (+6.6%), recall (+20.5%) and F1-score (+14.4%).

To enhance discriminative features at the label level K. Zhang et. al. [58] proposed LCFL method, which models both global and local correlations among labels, resulting in improved classification performance and state-of-the-art outcomes on the chexpert dataset. The proposed approach outperforms SOTA results with average F1 score (44.6%) and AUC (76.5%), which is 7.7% and 1.3% improvement, respectively. A self-attention is utilized to learn global relationship among labels by first extracting features at the image level

and then transforming them into label-specific representations. The model's performance is improved through consistency constraints and a contrastive loss framework designed for multi-label settings. The proposed LCFL method consists of two modules: LEFL and MLCL. To improve label-specific feature discrimination, LEFL focus on capturing local label relationships, while MLCL not only strengthens these correlations but also increases the distance between less related features. In the study, the following methods are used

- 1.Label relationships are modeled using a multi-head self-attention framework.
- 2.Global image features are broken down into label-wise feature representations.
- 3.A consistency constraint ensures alignment between global and local interactions.
- 4.A supervised contrastive loss tailored for multi-label settings is applied, leveraging label correlations.

The research by D. Mahapatra et.al.[77] introduced an approach called GANDALF, which combines sample selection with data augmentation in multi-label setting, depicting the disease labels as graph nodes and learning good inter-label relationships via graph attention transformer (GAT). The proposed approach GANDALF, addresses the challenge of generating informative samples, by sampling from a learned latent space, and of identifying informative samples through novel multi-label informativeness score. Designed for multi-label classification tasks, the proposed framework builds upon advanced active learning techniques by treating disease labels as nodes in a graph. It employs GAT to effectively learn and model relationships between labels. The method operates in two phases first, it selects the most informative data points using a multi-label evaluation metric and second it creates synthetic samples that add new informatino while avoiding redundancy.VAE generates images and baseline number of randomly selected training samples are used to train an initial classification model. Informative samples are selected by computing saliency maps for unlabeled data and building a graph that represents the underlying latent features extracted from class specific saliency maps.Informativeness score of each image is computed using a graph-based method. The aim is to represent each image as an individual graph, with nodes corresponding to potential class labels, and edge weights corresponding to the similarity between nodes. A graph transformer is proposed to learn to get more global relationships between graph nodes and to rank pool samples according to their informativeness. On the basis of the graph attention transformers, novel score termed as multi-label informativeness score, is derived.

Z. Lu et al. [78] suggested a framework MITER that focuses on overcoming the limitations of existing medical image-text pre-training models by leveraging large-scale generic datasets and small-scale medical datasets. It employs multi-level contrastive learning-based joint adaptive pretraining framework with two kinds of objectives: uni-modal objectives for pretraining image and text models separately, and the cross-modal objective for joint pre-training. The foundation of framework is using existing models which are pre-trained on general data and fine-tuned to medical domain through SSL. The proposed framework incorporates an adaptive hard negative sampling method, as well as a set of objectives, such as contrastive objectives and image-text matching objectives. To support unimodal contrastive learning, adaptive hard negative sampling is employed to form informative positive samples and difficult negative examples.

F Zheng, et al. [79] addressed the challenging problem caused by low resource level in medical field, highly professional annotation is required. This is a problem that can be mitigated by the use of a transferable representation of a large-scale pre-trained vision language model such as CLIP. In order to generate medical prompts auto-matically, MedPrompt, weakly-supervised prompt learning model is provided. The framework merges prompt learning under weak supervision with a vision-language model that has been pre-trained using unsupervised techniques. Using only the dataset's image categories, the weakly supervised prompt learning model independently learns the representation of each class-specific prompt vector.The proposed MedPrompt framework achieves superior performance on all four datasets (CheXpert, MIMIC-CXR, COVID and RSNA) in full-shot learning setup and shows superior accuracy when used for the zero-shot image classification and few-shot learning tasks. The study shows that Med- Prompt outperforms the previous models on CheXpert, MIMIC-CXR, and COVID datasets and is only slightly less effective than MedCLIP on RSNA dataset. The model outperforms the current models trained under fully-supervised setting on CheXpert, MIMIC-CXR and COVID in zero-shot classification. The model can correctly determine position of image with disease symptom and position of corresponding lung/lesion, and can correctly identify position of cardiac area and the area of the lung that has clear consolidation. The prompt generator is light weight, needs just a few additional parameters and computational overhead and can be embedded in any network architecture.

S J Park, et al. [80] presented medical X-VL, a vision-language framework based on cross attention mechanisms, designed with key adaptations for medical use. It supports zero-shot monitoring tasks and surpasses current state-of-art methods on two medical image benchmarks. Compared to current SOTA medical vision-language models, the medical X-VL model achieves improved performance at the tasks of zero-shot disease detection, and even zero-shot detection and correction of human errors. The model can show a finer-grained understanding of multiple abnormalities, such as errors in error detection, and the method facilitates zero-shot monitoring by identifying clinically relevant abnormalities and generating appropriate notifications. The proposed model demonstrates its superior performance for error detection

tasks comparing to current medical vision-language models, especially in error detection of location and extent. The model is credited to its capability of learning rich semantics and a wide range of visual concepts from free-form text. The proposed approach is a fusion encoder, which is a cross-attention between two uni-modal encoders. Trained visual encoder using the used DINO pre-trained model and self-supervised CXR-BERT model as the text encoder on the MIMIC- CXR training set. In this paper, they have presented a contrastive learning approach that is suited for medical domain data, including a contrastive learning method between every single sentence and image, to guarantee that individual observations are properly suited to visual presentation.

J.Rocha, et al (59) developed STERN, an automated system for detecting thoracic abnormalities in chest x-rays. It combines a CNN with an attention-based spatial transformer to focus on relevant thoracic regions, refine the input image and determine whether it is normal or abnormal. STERN method can produce outcomes that are comparable to those of cropping photographs utilizing YOLO, while also being less computationally expensive and not requiring localization labels. The system can differentiate between abnormal frontal images and CheXpert dataset with mean AUC (85.6%), which is 2.55% higher than baseline classifier. STERN approach achieved best performance across most of established metrics: AUC was increased by 2.55% over the single classifier. STERN loss turned out to be useful to fine tune, which led to improvement in the AUC of 0.22% in contrast to initial model. False positive (FP) predictions were reduced while increasing True Positive (TP) predictions using the STERN approach by removing misleading image artifacts. The STERN approach also gives better performance than the traditional approaches such as YOLO-based architectures, in terms of selecting the thoracic region of interest and classification of chest X-ray images as normal and abnormal with an improvement of 2.55% in AUC and 6.98% in precision. The four convolutional-layer ST module is more computation-ally efficient and can crop out more confounding features, such as letters, which are generally found at top of photos, when compared to the YOLO.

7. DISCUSSION

The 26 included studies span six methodological families and reveal a clear evolution of the field between 2020 and 2024. We synthesize the findings under five cross-cutting themes.

7.1 Architectural trends

Early CheXpert work was dominated by ImageNet pretrained CNN backbones—DenseNet-121, ResNet-50 and Xception—used either as plain classifiers or as feature extractors for downstream heads [73–75]. From 2022 on-wards a clear shift toward architectural specialization is visible: anatomy-aware attention [55], multi-scale residual fusion [66], dual-view fusion [69] and graph-convolutional inter-image reasoning [70]. The most recent work [78–80] adopts vision– language pre-training,

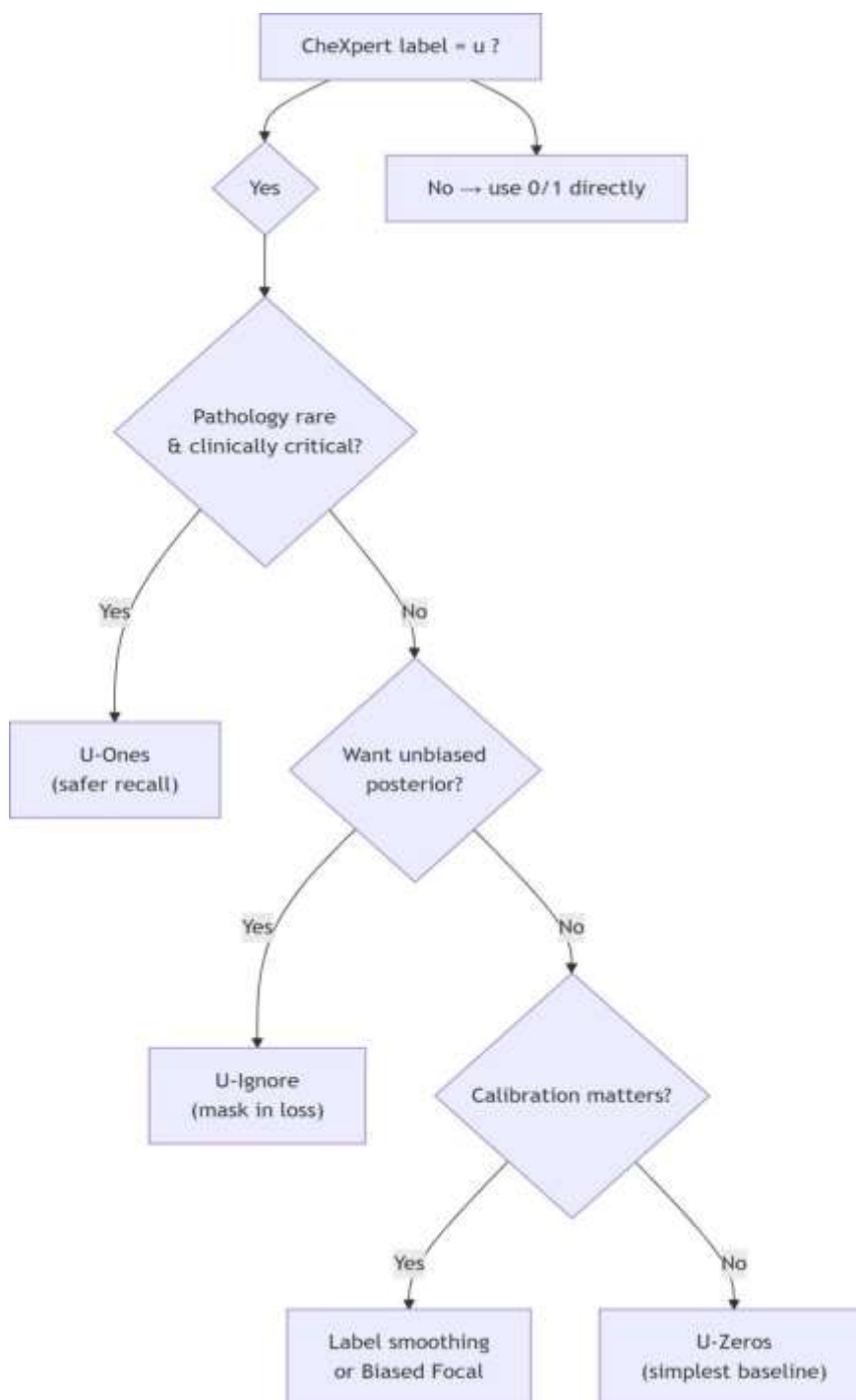


Fig. 7: Uncertainty Decision Tree

indicating that CheXpert is now used predominantly as a downstream benchmark for foundation model adaptation rather than as a stand-alone training corpus.

7.2 Handling of uncertain labels

Uncertainty handling remains the most CheXpert-specific design choice. As Table 4 shows, U-Ignore is the modal strategy among recent work, while early studies favored the simpler U-Zeros or U-Ones mappings. Methodological refinements—label smoothing [57], biased focal loss [66] and confidence-prediction heads [68]—consistently outperform naive mappings, but no study performs a systematic head-to-head comparison on a single backbone. This is a clear opportunity for future benchmarking work, building on the conditional-training scheme of Pham et al. [54].

7.3 Performance and reproducibility

Reported AUC values are difficult to compare directly: studies use heterogeneous validation splits (the 200-image expert-labelled validation set, the official hidden test set via the leaderboard, custom splits, or external datasets), and many report only mean AUC without per-pathology numbers or confidence intervals. Among the studies that report comparable five-pathology AUC, values cluster between 0.85 and 0.92, with diminishing returns from architectural complexity alone. Methods that incorporate multi-modal

signals [63, 76] or anatomical priors [55] show the largest relative gains over a plain DenseNet-121 baseline (in the spirit of CheXNet [81]). Less than half of the included studies make training code available, and even fewer make pre-trained weights available.

7.4 Robustness, calibration and explainability

Only one study [57] systematically evaluates robustness to common image perturbations across a large model zoo, and it finds that ImageNet accuracy is a better predictor of robustness than CheXpert AUC—a counter-intuitive result with direct implications for deployment. Calibration (e.g. ECE, Brier score) is essentially unreported. Explainability is addressed primarily through Grad-CAM [67] or attention visualization [55]; quantitative evaluation of explanations against radiologist annotations is rare.

7.5 Bias, fairness and external generalizability

None of the included studies performs a stratified fairness analysis along sex, age or self-reported race, despite published evidence [82] that CheXpert-trained models systematically under-diagnose under-served populations. External validation on independent datasets such as MIMIC-CXR [25], PadChest [26] or VinDr-CXR [29] is performed in fewer than one third of the studies, and reported drops of 5–15 AUC points on out-of-distribution data are typical. This represents the single largest gap between the methodological literature and clinical readiness.

7.6 Limitations of this review

This review is restricted to English-language work and to studies in which CheXpert is a primary dataset; methodological papers that use CheXpert only for ablations may have been missed. Reported AUC values are taken as published and have not been independently re-computed; differences in validation splits limit direct comparability. Finally, the rapidly growing foundation-model literature means that the cut-off date of March 2026 may already omit relevant pre-prints.

8. OPEN CHALLENGES AND FUTURE DIRECTIONS

8.1 Foundation models and parameter-efficient adaptation

The emergence of medical vision-language and self-supervised foundation models—CheXzero [60], BiomedCLIP [83], RAD-DINO [84] and CheXagent [85]—together with

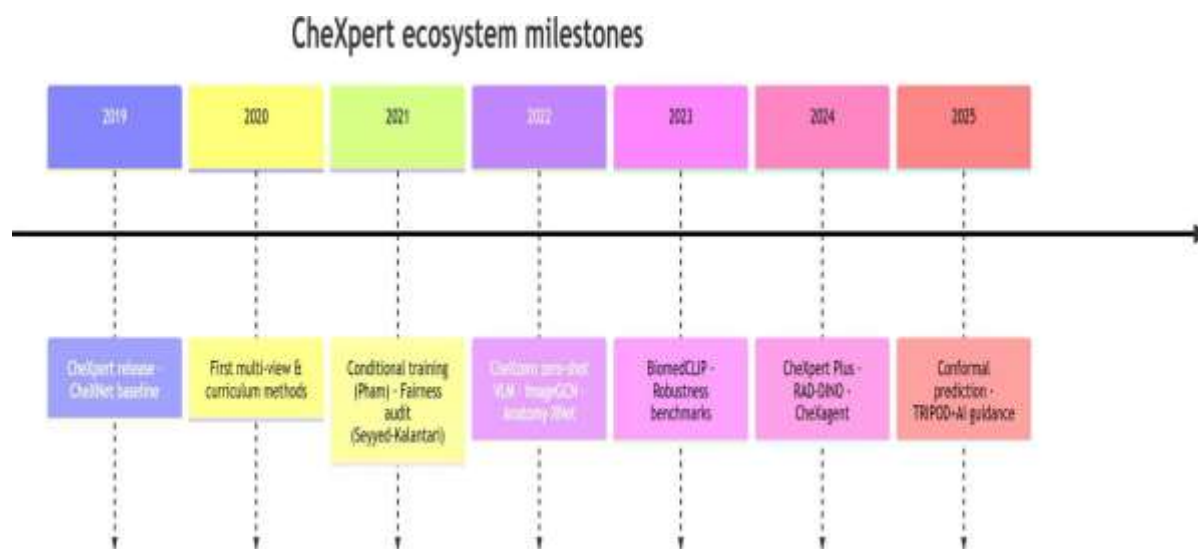


Fig. 8: Milestones timeline

the release of the augmented CheXpert Plus corpus [86] shifts the question from “which CNN backbone” to “which adaptation strategy”. LoRA, prompt tuning and adapter modules deserve a systematic comparison on CheXpert.

8.2 Federated and privacy-preserving training

Multi-institutional federated learning across CheXpert, MIMIC-CXR and PadChest could address the generalization gap without raw-data sharing; recent benchmarks [87] show that the methodology is now mature, but differential-privacy guarantees on CheXpert remain largely unexplored.

8.3 Calibration-aware learning

Reliable triage requires calibrated probabilities, not just discriminative AUC. Temperature scaling, focal calibration and conformal prediction [88] should become reporting defaults.

8.4 Fairness auditing as a standard reporting item

Subgroup AUC and equalised-odds gaps along age, sex and race should accompany any new CheXpert benchmark result, mirroring the recommendations of CLAIM and the updated TRIPOD+AI statement [89].

8.5 Standardized evaluation protocol

A community-agreed protocol—fixed validation split, mandatory five- and fourteen-pathology reporting.

per-pathology 95% confidence intervals, and at least one external test set—would dramatically improve the comparability that this review repeatedly found lacking.

8.6 Prospective clinical validation

Beyond retrospective AUC, reader studies and silent-mode prospective deployments are required to demonstrate decision-impact and workflow benefit.

8.7 Uncertainty quantification

Beyond the dataset's own label uncertainty, predictive uncertainty (Bayesian, ensemble, evidential) is essential for safe deferral to a radiologist; current CheXpert work largely ignores it.

9. SUMMARY AND CONCLUSIONS

We have presented a PRISMA-guided review of 26 deep learning studies that target the CheXpert dataset for the classification of pulmonary disease. The reviewed work traces a clear arc from supervised CNN baselines, through architectural specialization (anatomy-aware, multi-scale, multi-view and graph-based models), to the current generation of self-supervised and vision-language approaches. Across this trajectory, three observations are robust:

1. Handling of CheXpert's uncertainty labels remains a decisive design choice, with the U-Ignore strategy and its principled extensions consistently outperforming naive mappings.
2. Architectural innovation alone yields diminishing returns, and the largest gains now come from multi-modal signals, anatomical priors and large-scale pre-training.
3. Methodological progress has outpaced the rigour of evaluation, with limited external validation, almost no calibration or fairness reporting, and inconsistent code release.

For the field to translate these advances into clinical value, the community needs standardized evaluation protocols, mandatory subgroup and calibration reporting, federated multi-institution benchmarks, and prospective clinical studies. We hope that the taxonomy, comparison tables and open challenges synthesized in this review will help researchers and clinicians position new contributions within a coherent landscape and accelerate the responsible deployment of deep learning for chest-radiograph interpretation.

Supplementary information. : 'Not applicable'

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

• **Funding:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

• **Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use):** 'Not Applicable'

• **Ethics approval and consent to participate:** This work is a literature review and did not involve human participants, human data, or animals; institutional ethics review was therefore not required.

•Consent for publication

• **Data availability :** No new data were generated for this review. All primary studies analyzed are publicly indexed and cited. The CheXpert dataset is available from the Stanford Machine Learning Group at <https://stanfordmlgroup.github.io/competitions/chexpert/>.

• **Materials availability:** 'Not applicable'

• **Code availability :** 'Not applicable'

• **Author contribution:** Kirti Swagat Mohanty: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Visualization. Chandrajit Choudhury: Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. Prabina Pattanayak: Methodology, Validation, Writing – review & editing, Supervision, Project administration.

References

1. Shaw, N.J., Hendry, M., Eden, O.B.: Inter-observer variation in interpretation of chest x-rays. *Scottish Medical Journal* 35(5), 140–141 (1990)
2. Yasaka, K., Abe, O.: Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLOS Medicine* 15 (2018)
3. Wang, H., Nagavi, M., et al.: Global, regional and national life expectancy, all-cause mortality, and cause-specific mortality for 249 cases of death, 1980–2015: A systematic analysis for the global burden of disease study 2015. *The Lancet* 388(10053) (2016)
4. Wielputz, M.O., Heubel, C.P., Herth, F.J.F., Kauczor, H.-U.: Radiological diagnosis in lung disease. *Dtsch Arztebl International* 111(11), 181–187 (2014) <https://doi.org/10.3238/arztebl.2014.0181>
5. Wu, C., Luo, C., Xiong, N., Zhang, W., Kim, T.-H.: A greedy deep learning method for medical disease analysis. *IEEE Access* 6, 20021–20030 (2018)
6. Ma, J., Song, Y., Tian, X., Hua, Y., Zhang, R., Wu, J.: Survey on deep learning for pulmonary medical imaging. *Frontiers of Medicine* 14(4), 450–469 (2020)

7. Kieu, S.T.H., Bade, A., Hijazi, M.H.A., Kolivand, H.: A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions. *Journal of Imaging* 6(12), 131 (2020)
8. C, alli, E., Sogancioglu, E., Ginneken, B., Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis* 72, 102125 (2021)
9. Meedeniya, D., Kumarasinghe, H., Kolonne, S., Fernando, C., Torre D'íez, I., Marques, G.: Chest x-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing* 126, 109319 (2022)
10. Singh, R., Kalra, M.K., Nitiwarangkul, C., Patti, J.A., Homayounieh, F., Padole, A., Rao, P., Putha, P., Muse, V.V., Sharma, A., et al.: Deep learning in chest radiography: detection of findings and presence of change. *PloS One* 13(10), 0204155 (2018)
11. Qin, C., Yao, D., Shi, Y., Song, Z.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomedical Engineering Online* 17, 1–23 (2018)
12. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., Pfeiffer, D.: Efficient deep net-work architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific Reports* 9(1), 6268 (2019)
13. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., S'anchez, C.I.: A survey on deep learning in medical image analysis. *Medical Image Analysis* 42, 60–88 (2017)
14. Feng, Y., Teh, H.S., Cai, Y.: Deep learning for chest radiology: a review. *Current Radiology Reports* 7, 1–9 (2019)
15. Ginneken, B.: Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology* 10, 23–32 (2017)
16. Sahiner, B., Pezeshk, A., Hadjiiski, L.M., Wang, X., Drukker, K., Cha, K.H., Summers, R.M., Giger, M.L.: Deep learning in medical imaging and radiation therapy. *Medical Physics* 46(1), 1–36 (2019)
17. Kallianos, K., Mongan, J., Antani, S., Henry, T., Taylor, A., Abuya, J., Kohli, M.: How far have we come? artificial intelligence for chest radiograph interpretation. *Clinical Radiology* 74(5), 338–345 (2019)
18. Anis, S., Lai, K.W., Chuah, J.H., Ali, S.M., Mohafez, H., Hadizadeh, M., Yan, D., Ong, Z.-C.: An overview of deep learning approaches in chest radiograph. *IEEE Access* 8, 182347–182354 (2020)
19. Khan, W., Zaki, N., Ali, L.: Intelligent pneumonia identification from chest x-rays: A systematic literature review. *IEEE Access* 9, 51747–51771 (2021)
20. Pratim, D., Nachamai, M.: A review on deep learning method for lung cancer stage classification using pet-ct. In: *Applied Smart Health Care Informatics: A Computational Intelligence Perspective*, pp. 9–29. Springer, ??? (2022)
21. Oloko-Oba, M., Viriri, S.: A systematic review of deep learning techniques for tuberculosis detection from chest radiograph. *Frontiers in Medicine* 9, 830515 (2022)
22. Li, R., Xiao, C., Huang, Y., Hassan, H., Huang, B.: Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review. *Diagnostics* 12(2), 298 (2022)
23. Chen, Y., Jiang, G., Li, Y., Tang, Y., Xu, Y., Ding, S., Xin, Y., Lu, Y.: A survey on artificial intelligence in chest imaging of covid-19. *BIO Integration* 1, 137–146 (2020)
24. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597 (2019)
25. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-Y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG: a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019)
26. Bustos, A., Pertusa, A., Salinas, J.-M., De La Iglesia-Vaya, M.: PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 66, 101797 (2020)
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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* (2017)
28. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23(2), 304–310 (2016)
29. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.B., Pham, C.M., Tong, H.T.T., Dinh, D.H., et al.: VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data* 9(1), 429 (2022)
30. Zhu, C.S., Pinsky, P.F., Kramer, B.S., Prorok, P.C., Purdue, M.P., Berg, C.D., Gohagan, J.K.: The prostate, lung, colorectal, and ovarian cancer screening trial and its associated research resource. *Journal of the National Cancer Institute* 105(22), 1684–1693 (2013)
31. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* 160(1), 106 (1962)

32. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36(4), 193–202 (1980)
33. Rumelhart, D., Hinton, G., Williams, R.: Learning representations by back-propagation errors. *Nature* 323, 533–536 (1986)
34. Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: *Neural Networks for Perception*, pp. 65–93. Elsevier, ??? (1992)
35. LeCun, Y.: Generalization and network design strategies. *Connections in Perspective* (1989)
36. Hinton, G.E., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. *Neural Computation* 18(7), 1527–1554 (2006)
37. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25 (2012)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)
40. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
41. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708 (2017)
42. Krogh, A., Hertz, J.: Dynamics of generalization in linear perceptrons. In: *Advances in Neural Information Processing Systems*, vol. 3 (1990)
43. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. In: *Advances in Neural Information Processing Systems*, vol. 2 (1989)
44. Haskell, B.G., Howard, P.G., LeCun, Y.A., Puri, A., Ostermann, J., Civanlar, M.R., Rabiner, L., Bottou, L., Haffner, P.: Image and video coding—emerging standards and beyond. *IEEE Transactions on Circuits and Systems for Video Technology* 8(7), 814–837 (1998)
45. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *Technical report* (2001)
46. Divina, F., Gilson, A., Gómez-Vela, F., García Torres, M., Torres, J.F.: Stacking ensemble learning for short-term electricity consumption forecasting. *Energies* 11(4), 949 (2018)
47. Lincoln, W., Skrzypek, J.: Synergy of clustering multiple backpropagation networks. In: *Advances in Neural Information Processing Systems*, vol. 2, pp. 650–659 (1989)
48. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis* 58, 101539 (2019)
49. Huang, S.-C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine* 6(1), 74 (2023)
50. Rajaraman, S., Cemir, S., Xue, Z., Alderson, P., Thoma, G., Antani, S.: A novel stacked model ensemble for improved tb detection in chest radiographs. In: *Medical Imaging*, pp. 1–26. CRC Press, ??? (2019)
51. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* 25(6), 954–961 (2019)
52. Gordienko, Y., Gang, P., Hui, J., Zeng, W., Kochura, Y., Alienin, O., Rokovyi, O., Stirenko, S.: Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer. In: *Advances in Computer Science for Engineering and Education* 13, pp. 638–647. Springer, ??? (2019)
53. Hijazi, M.H.A., Hwa, S.K.T., Bade, A., Yaakob, R., Jeffree, M.S.: Ensemble deep learning for tuberculosis detection using chest x-ray and canny edge detected images. *IAES International Journal of Artificial Intelligence* 8(4), 429 (2019)
54. Pham, H.H., Le, T.T., Tran, D.Q., Nguyen, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing* 437, 186–194 (2021)
55. Kamal, U., Zunaed, M., Nizam, N.B., Hasan, T.: Anatomy-XNet: An anatomy aware convolutional neural network for thoracic disease classification in chest x-rays. *IEEE Journal of Biomedical and Health Informatics* 26(11), 5518–5528 (2022)
56. Termritthikun, C., Umer, A., Suwanwimolkul, S., Xia, F., Lee, I.: Explainable knowledge distillation for on-device chest x-ray classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2023)
57. Chong, C.F., Fang, X., Yang, X., Luo, W., Wang, Y.: On the correlations between performance of deep

- networks and its robustness to common image perturbations in medical image interpretation. In: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 426–433. IEEE, ??? (2023)
58. Zhang, K., Liang, W., Cao, P., Liu, X., Yang, J., Zaiane, O.: Label correlation guided discriminative label feature learning for multi-label chest image classification. *Computer Methods and Programs in Biomedicine* 245, 108032 (2024)
59. Rocha, J., Pereira, S.C., Pedrosa, J., Campilho, A., Mendonça, A.M.: STERN: Attention-driven spatial transformer network for abnormality detection in chest x-ray images. *Artificial Intelligence in Medicine* 147, 102737 (2024)
60. Tiu, E., Talus, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering* 6(12), 1399–1406 (2022)
61. Zhou, H.-Y., Lu, C., Chen, C., Yang, S., Yu, Y.: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(7), 8020–8035 (2023)
62. Costa, M., Pereira, S.C., Pedrosa, J., Mendonça, A.M., Campilho, A.: Deep feature-based automated chest radiography compliance assessment. In: 2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG), pp. 64–67. IEEE, ??? (2023)
63. Meshnick, D.C., Shahini, N., Ganguly, D., Wu, Y., French, R.H., Chaudhary, V.: Enhancing scientific image classification through multimodal learning: Insights from chest x-ray and atomic force microscopy datasets. In: 2023 IEEE International Conference on Big Data (BigData), pp. 2211–2220. IEEE, ??? (2023)
64. Nie, W., Zhang, C., Song, D., Bai, Y., Xie, K., Liu, A.: Instrumental variable learning for chest x-ray classification. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4506–4512. IEEE, ??? (2023)
65. Liu, Z., Cheng, Y., Tamura, S.: Multi-label local to global learning: a novel learning paradigm for chest x-ray abnormality classification. *IEEE Journal of Biomedical and Health Informatics* 27(9), 4409–4420 (2023)
66. Li, Q., Lai, Y., Adamu, M.J., Qu, L., Nie, J., Nie, W.: Multi-level residual feature fusion network for thoracic disease classification in chest x-ray images. *IEEE Access* 11, 40988–41002 (2023)
67. Talasila, A., Karthikeyan, A., Alle, S., Maity, M., Priyakumar, U.D.: Self-supervision and weak supervision for accurate and interpretable chest x-ray classification models. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, ??? (2023)
68. Rabbani, N., Bartoli, A.: Unsupervised confidence approximation: Trustworthy learning from noisy labelled data. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4609–4617 (2023)
69. Ni, J., Bhuyan, Z., Chen, Q., Sun, X., Wang, D., Cao, Y., Liu, B.: Enhance chest x-ray classification with multi-image fusion and pseudo-3d reconstruction. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, ??? (2022)
70. Mao, C., Yao, L., Luo, Y.: ImageGCN: Multi-relational image graph convolutional networks for disease identification with chest x-rays. *IEEE Transactions on Medical Imaging* 41(8), 1990–2003 (2022)
71. Biesner, D., Schneider, H., Wulff, B., Attenberger, U., Sifa, R.: Improving chest x-ray classification by rnn-based patient monitoring. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 946–950. IEEE, ??? (2022)
72. Wang, Z., Zhang, X., Chen, W., Niu, J.: Lung segmentation reconstruction based data augmentation approach for abnormal chest x-ray images diagnosis. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2203–2207. IEEE, ??? (2022)
73. Chen, W., Feng, J., Zhou, J.: TriageNet: A multi-agent diagnosis network for imbalanced data. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE, ??? (2022)
74. Crespi, L., Loiacono, D., Chiti, A.: Chest x-rays image classification from variational autoencoders latent features. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8. IEEE, ??? (2021)
75. Mitra, A., Chakravarty, A., Ghosh, N., Sarkar, T., Sethuraman, R., Sheet, D.: A systematic search over deep convolutional neural network architectures for screening chest radiographs. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1225–1228. IEEE, ??? (2020)
76. Ren, H., Jing, F., Chen, Z., He, S., Zhou, J., Liu, L., Jing, R., Lian, W., Tian, J., Zhang, Q., et al.: ChexMed: A multimodal learning algorithm for pneumonia detection in the elderly. *Information Sciences* 654, 119854 (2024)
77. Mahapatra, D., Bozorgtabar, B., Ge, Z., Reyes, M.: GANDALF: Graph-based transformer and data augmentation active learning framework with interpretable features for multi-label chest xray classification. *Medical Image Analysis* 93, 103075 (2024)

78. Shu, C., Zhu, Y., Tang, X., Xiao, J., Chen, Y., Li, X., Zhang, Q., Lu, Z.: MITER: Medical image-text joint adaptive pretraining with multi-level contrastive learning. *Expert Systems with Applications* 238, 121526 (2024)
79. Zheng, F., Cao, J., Yu, W., Chen, Z., Xiao, N., Lu, Y.: Exploring low-resource medical image classification with weakly supervised prompt learning. *Pattern Recognition* 149, 110250 (2024)
80. Park, S., Lee, E.S., Shin, K.S., Lee, J.E., Ye, J.C.: Self-supervised multi-modal training from uncured images and reports enables monitoring ai in radiology. *Medical Image Analysis* 91, 103021 (2024)
81. Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P.: Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(2), 652–662 (2019)
82. 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.: CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017)
83. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radio-graphs in under-served patient populations. *Nature Medicine* 27(12), 2176–2182 (2021)
84. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: BiomedCLIP: a multimodal biomedical foundation model pre-trained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023)
85. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: RAD-DINO: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815* (2024)
86. Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: CheXa-gent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208* (2024)
87. Chambon, P., Delbrouck, J.-B., Sounack, T., Huang, S.-C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics, and additional image formats. *arXiv preprint arXiv:2405.19538* (2024)
88. Ng, D., Lan, X., Yao, M.M.-S., Chan, W.P., Feng, M.: Federated learning for chest x-ray classification: A systematic benchmark across institutions. *npj Digital Medicine* 7(1), 89 (2024)
89. Lu, C., Ahmed, F., Singh, R., Lungren, M.P., Rajpurkar, P.: Conformal pre-diction for reliable diagnosis from medical images: A systematic study on chest radiographs. *Medical Image Analysis* 99, 103352 (2025)