Enhancing CheXNet with Uncertainty Quantification

Project ID:- MED002

Index No:- 210329E
Name:- Laksara K.Y

1. Introduction and Project Framework

1.1. Problem Statement

Machine learning models show impressive results in research, but their use in high-stakes clinical practice is limited. A key reason is the absence of reliable confidence measures [1].Traditional deep learning models provide only a single point prediction without an accompanying quantification of its trustworthiness. This is particularly problematic in medical image analysis, where an incorrect or overconfident prediction can have severe consequences [2]. Even models with superhuman performance often fail clinical deployment because clinicians lack actionable trust in predictions.This project specifically bridges that gap [3].

CheXNet, a landmark deep learning model based on DenseNet-121, demonstrated radiologist-level performance in detecting pneumonia on the NIH ChestX-ray14 dataset and has since become a state-of-the-art benchmark in thoracic disease detection. However, its predictions remain purely probabilistic and do not provide calibrated or uncertainty-aware measures of trustworthiness. This gap limits its clinical applicability, as practitioners require not only accurate predictions but also an understanding of how confident the model is in those predictions [4].

1.2. Project Objectives

- Implement Monte Carlo Dropout (MCD) in the CheXNet model to quantify prediction uncertainty.
- Evaluate MCD in terms of model reliability, calibration, and classification performance (AUROC, F1-score).
- Ensure MCD-enhanced CheXNet meets or exceeds the original CheXNet SOTA baseline.
- Explore Deep Ensembles (DE) only if MCD performance is insufficient, to further improve uncertainty estimation.
- Demonstrate measurable performance gains in model trustworthiness and calibration, not just raw classification metrics.
- Prepare and submit a conference format research paper detailing methodology, results, limitations, and clinical implications.

2. CheXNet (SOTA Model) and Benchmark Datasets

2.1. CheXNet Model

The CheXNet model, a landmark state-of-the-art (SOTA) model introduced in a 2017 paper by a Stanford research team, used a 121-layer Dense Convolutional Network (DenseNet) to detect pneumonia from chest X-rays. The model achieved state-of-the-art results on all 14 pathology classes and demonstrated performance exceeding the average practicing radiologist on the F1 metric for pneumonia detection.The original paper reported that CheXNet's per-class AUROC scores on the test set outperformed previous SOTA results by a margin of >0.05 AUROC on Mass, Nodule, Pneumonia, and Emphysema. The network uses ImageNet-pretrained weights and is trained end-to-end with the Adam optimizer (β_1=0.9, β_2=0.999), a batch size of 16, and an initial learning rate of 0.0001 that is decayed when the validation loss plateaus. The final fully connected layer is replaced with 14 outputs (one per pathology) with a sigmoid activation, enabling multi-label classification [4].

Table 1: AUROC performance of CheXNet versus prior state-of-the-art models

Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (ours)
Atelectasis	0.716	0.772	**0.8094**
Cardiomegaly	0.807	0.904	**0.9248**
Effusion	0.784	0.859	**0.8638**
Infiltration	0.609	0.695	**0.7345**
Mass	0.706	0.792	**0.8676**
Nodule	0.671	0.717	**0.7802**
Pneumonia	0.633	0.713	**0.7680**
Pneumothorax	0.806	0.841	**0.8887**
Consolidation	0.708	0.788	**0.7901**
Edema	0.835	0.882	**0.8878**
Emphysema	0.815	0.829	**0.9371**
Fibrosis	0.769	0.767	**0.8047**
Pleural Thickening	0.708	0.765	**0.8062**
Hernia	0.767	0.914	**0.9164**

2.2. NIH ChestX-ray14 Dataset (Primary Benchmark Dataset)

The NIH ChestX-ray14 dataset is a large-scale collection of frontal-view chest X-ray images from 30,805 unique patients, totaling 112,120 images. Each image is annotated with text-mined labels for fourteen common thoracic diseases, including Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia.[9] These labels were extracted from associated radiological reports using natural language processing (NLP), with an expected accuracy above 90% [5].

The dataset is significantly larger than previous public chest X-ray datasets, making it more representative of real patient populations and providing a richer resource for deep learning-based medical imaging research [5]. Despite its size, the dataset exhibits severe class imbalance, with rare diseases like Hernia represented by very few images, while common conditions such as Infiltration are much more prevalent [6]. Additionally, many images contain multiple disease labels, reflecting realistic co-occurrence patterns.

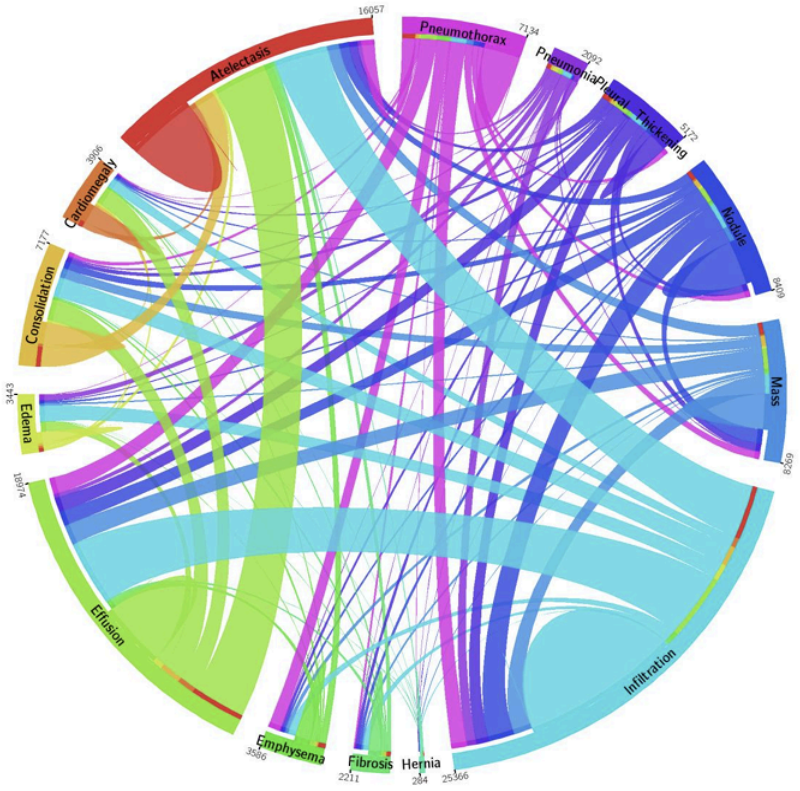

Figure 1: Distributions of 14 disease categories with co-occurrence statistics

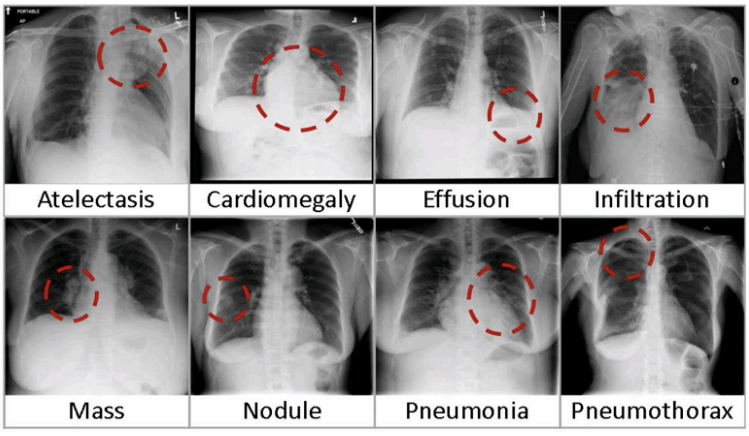

Figure 2: Eight visual examples of common thorax diseases.

The dataset is divided at the patient level into training/validation and test sets to prevent patient overlap, ensuring unbiased evaluation [5]. This dataset provides a robust benchmark for model development, particularly for multi-label classification tasks and methods addressing class imbalance.

2.3. Additional Benchmark Datasets

While the NIH ChestX-ray14 dataset is the primary dataset for training and model selection, performance may also be referenced against standard benchmarks such as CheXpert and MIMIC-CXR to situate our results within the broader literature. The CheXpert dataset contains 224,316 images from 65,240 patients and is a notable benchmark that includes uncertainty labels [7]. The MIMIC-CXR dataset is even larger, consisting of 473,064 chest X-rays with associated radiology reports from 63,478 patients, making it over four times the size of ChestX-ray14 [8]. This dataset is known for its high-resolution images and is a crucial resource for comparing AI models [9].

3. Literature Review

3.1. Importance of Uncertainty in Medical AI

As established in the introduction, quantifying uncertainty is crucial for the clinical adoption of AI models in medical imaging, as it provides a necessary measure of confidence that goes beyond raw accuracy. A model's total uncertainty can be broken down into two main types:

- **Aleatoric Uncertainty:** This is the inherent, irreducible noise in the data itself. In medical imaging, this can be caused by factors like sensor noise, imaging artifacts, or, critically, ambiguous labels from the source radiology reports, which often use phrases like "could be due to" or "cannot be excluded." A model cannot reduce this type of uncertainty by seeing more data [10].
- **Epistemic Uncertainty:** This is the model's own uncertainty due to a lack of knowledge or limited data. It can be reduced by providing the model with more data and is often high for out-of-distribution (OOD) or "never-before-seen" samples [10] .

A key finding from recent research is that incorporating uncertainty labels during model training leads to higher predictive variance for uncertain cases at test time, preventing the model from making "over-confident mistakes" [10]. This ability is of significant clinical value, particularly when a model flags an out-of-distribution case that a clinician should review. The field of UQ for medical imaging is a growing area of research, with recent reviews providing a comprehensive overview of both probabilistic and non-probabilistic methods.

3.2. Related Work

The field of Uncertainty Quantification in deep learning has seen a variety of approaches, spanning both probabilistic and non-probabilistic methods, particularly in medical image analysis. My work builds upon a foundation of extensive benchmarking and exploration of these methods [11].

These studies often categorize methods into "distributional" and "deterministic" approaches. Distributional methods, such as Latent Heteroscedastic Classifiers, model uncertainty by learning a second-order predictive distribution, where the output is a distribution over predictions rather than a single point estimate. This category includes a wide array of techniques, such as SWAG, Evidential Deep Learning (EDL), and the Deep and Shallow Ensembles I am exploring [11].

In contrast, deterministic methods, like Loss Prediction or Temperature Scaling, estimate uncertainty without requiring a predictive distribution, often relying on model-internal features to assess confidence. Other research has also explored the information-theoretical approach for disentangling epistemic and aleatoric uncertainty, though recent studies have raised concerns about its practical effectiveness in complex, real world datasets like those used in medical imaging [12].

3.3. Methods for Uncertainty Quantification

Several uncertainty quantification approaches exist in the literature. For clarity, this section first provides a broad overview of both probabilistic and non-probabilistic methods. This project, however, focuses specifically on Monte Carlo Dropout and Deep Ensembles.

Uncertainty quantification methods can be broadly categorized into probabilistic and non-probabilistic approaches, which can be further sub-divided into various techniques. Probabilistic methods, such as Bayesian deep learning, model uncertainty by learning a full predictive distribution, while non-probabilistic methods rely on alternative heuristics to estimate confidence. The choice of method often depends on the specific application, computational budget, and desired level of theoretical rigor.

Probabilistic Methods: Bayesian Deep Learning and Approximations

Bayesian deep learning (BDL) is a promising approach for uncertainty quantification in healthcare as it provides a probabilistic framework that quantifies uncertainty and enhances prediction reliability. This approach has proven advantageous in applications requiring nuanced decision making, such as early disease detection and personalized treatment plans [13]. While the original CheXNet model does not include UQ, several Bayesian methods can be adapted for this purpose, with varying trade-offs in computational complexity and performance.

Directly implementing a full Bayesian Neural Network (BNN) using methods like Markov Chain Monte Carlo (MCMC) is often computationally intractable for large, modern neural networks due to the high-dimensional parameter space. MCMC methods, while theoretically sound, produce autocorrelated samples and can be inefficient in exploring the target distribution, leading to challenges with convergence

and high computational costs. Therefore, researchers have developed various approximation methods that offer a balance between theoretical rigor and practical feasibility.

- **Variational Inference (VI):** This method approximates the true posterior distribution of the model's weights with a simpler, more tractable distribution, such as a Gaussian distribution. The goal is to make the approximate distribution as close as possible to the true posterior by minimizing a measure like the Kullback–Leibler (KL) divergence. While VI offers a principled Bayesian framework, it still has limitations and may not fully capture the complex uncertainty landscape [14].
- **Monte Carlo Dropout (MCD):** This is a practical and computationally efficient approximation of a BNN that leverages dropout, a technique traditionally used for regularization, by keeping it active during inference. For a single input image, multiple forward passes are performed with different neurons randomly "dropped out," effectively creating an ensemble of "random subnets" [15] . The final prediction is the average of these passes, and the uncertainty is quantified by the variance of the predictions. While more efficient than full Bayesian methods, MCD tends to sample from a single, slightly varied region of the loss landscape, yielding a "distribution of similar functions" [16],[17].
- **Deep Ensembles (DE):** Deep Ensembles offer a more robust, albeit more computationally intensive, approach to UQ that is often considered a non-Bayesian approach, although some have explored its links to Bayesian methods. The method involves training a collection of identical models with different random initializations. The diversity in initial weight values and training trajectories causes the models to converge to different, well separated "low-loss valleys" in the parameter space, leading to a "distribution of diverse functions" [18]. This diversity makes Deep Ensembles a state-of-the-art method for uncertainty quantification, particularly in out-of-distribution settings, where they have been shown to outperform MCD and other methods [18]. However, the computational cost of training multiple full models is a significant limitation, which researchers are trying to address with more efficient ensemble methods like Stochastic Weight Averaging in Parallel (SWAP) [19].
- **Ensemble Bayesian Neural Networks (EBNN):** This is a specific type of ensemble rooted in Bayesian theory. EBNN addresses epistemic uncertainty by learning a posterior distribution over the model's parameters. It involves sampling and weighting networks according to this posterior to form an ensemble model, also referred to as a "Bayes ensemble." This is conceptually distinct from a standard Deep Ensemble, as it leverages the probabilistic framework of BNNs. While theoretically sound, EBNN is often computationally demanding due to the complexity of sampling from the posterior distribution, and its performance may not always surpass simpler uniformly-weighted deep ensembles, according to recent research [17], [20].
- **Ensemble Monte Carlo (EMC) Dropout:** This is an extension of Monte Carlo Dropout designed to overcome its limitations. One recent approach proposes a strategy to compute an ensemble of subnetworks, each corresponding to a non-overlapping dropout mask, and trains them independently via a pruning strategy. The goal of this method is to bridge the performance gap between MC Dropout and Deep Ensembles, achieving similar accuracy and uncertainty estimates to deep ensembles while maintaining the computational efficiency of MC Dropout [21].

Non-Probabilistic Methods

In contrast to probabilistic methods that model a predictive distribution, non-probabilistic or deterministic methods estimate uncertainty based on internal model properties without a rigorous statistical framework. These approaches, often post-processing steps or architectural modifications, are typically computationally efficient but may provide less nuanced uncertainty estimates.

- **Conformal Prediction (CP):** This is a prominent non-probabilistic method that provides statistically rigorous prediction sets for each prediction. CP guarantees that the true label will be included in the prediction set at a user-specified error rate on average across the entire test distribution. However, this guarantee is marginal and does not hold for specific subgroups or individual data points, which can be a significant limitation in clinical applications where rare but critical classes may be systematically under-covered [22].
- **Temperature Scaling:** This is a post-hoc calibration technique that adjusts a model's confidence scores by dividing the logits (model outputs before the softmax layer) by a single scalar value called the "temperature." This method has been shown to improve the calibration of a model's confidence but is limited as the temperature is calculated on a validation set and may not generalize well to out-of-distribution data [23],[24].

For this project, only Monte Carlo Dropout (MCD) and Deep Ensembles (DE) will be implemented. These methods were chosen due to their proven effectiveness in multi-label medical image classification, relative computational feasibility, and ability to provide both epistemic and aleatoric uncertainty estimates. Other methods discussed are acknowledged for completeness but are beyond the scope of this work.

4. Proposed Methodology

4.1. Data Preprocessing and Augmentation

Given the variability in chest X-ray acquisition across hospitals, robust preprocessing and augmentation strategies will be applied. It is important to note that these augmentations will only be applied to the training dataset to ensure the validation and test sets remain pristine for an unbiased evaluation.

- **Resizing & Normalization:** All images resized to 224×224 and normalized to ImageNet statistics.
- **Data Augmentation:**
  - Horizontal flipping
  - Random rotations (±15°)
  - Contrast and brightness adjustments
  - Random zoom and cropping
  - Random translations and perspective distortions

These augmentations simulate real-world imaging variability and are expected to improve generalization

and reduce overfitting [25].

4.2. Addressing Class Imbalance

The ChestX-ray14 dataset suffers from extreme class imbalance (e.g., Hernia has <1% prevalence). To mitigate this, two strategies will be explored:

- **Loss Function Modifications:**
  - The original CheXNet paper used unweighted Binary Cross-Entropy (BCE) loss [26]. To address the class imbalance, I will explore weighted BCE and Focal Loss[27],[28] to down weight easy negatives and focus on rare classes.
- **Resampling Approaches:**
  - Oversampling of minority classes, cautiously applied to avoid overfitting.

The effectiveness of these strategies will be empirically validated to determine the most suitable approach.

4.3. Model Architecture and Enhancements

The baseline for this project will be a reproduced CheXNet model, which is a DenseNet-121 pre-trained on ImageNet with its final layer adapted for multi-label classification of the 14 thoracic diseases. The model will be trained using the same hyperparameters as in the original CheXNet implementation, and the pretrained weights will be fine-tuned on the ChestX-ray14 dataset to ensure reproducible baseline performance.

I plan to explore two distinct uncertainty quantification methods, starting with the Monte Carlo Dropout (MCD) approach due to its lower computational cost and relative ease of integration. If this initial approach does not yield a performance at or above the state-of-the-art (SOTA) benchmark on key metrics like AUROC, I will then proceed to explore the more computationally intensive Deep Ensembles method.

4.3.1 Adapting the Model for Monte Carlo Dropout

To implement Monte Carlo Dropout (MCDO), the standard CheXNet DenseNet-121 model must be modified to include Dropout layers. This is a crucial architectural enhancement, as the original CheXNet implementation and its replicas do not mention the use of dropout [27]. The model will be trained with dropout enabled, as is standard practice for regularization. The primary difference will occur during inference, where the model.train() function will be called to keep the dropout layers active, followed by multiple forward passes on a single input. For this project, I will start with T=20 as a reasonable tradeoff between computational cost and uncertainty resolution, but this will be tuned experimentally. The final class probability will be the mean of the T predictions, while the standard deviation or entropy of the predictions will serve as the estimate of epistemic uncertainty.

4.3.2 Design of a Deep Ensemble-Based CheXNet

For the Deep Ensemble approach, an ensemble of K independent DenseNet-121 models will be created. Each model will share the same architectural and training parameters as the baseline CheXNet and will be initialized with ImageNet pretrained weights, while allowing independent randomization of the final layers. For this project, the ensemble size will be explored in the range K = 2–5 models, allowing flexibility to balance computational cost and uncertainty estimation performance. This process is computationally intensive, as it requires training K full models [29]. For inference, each of the K models will produce a prediction, and the final class probability will be the mean of the K predictions. The variance among the K predictions will be used to quantify both aleatoric and epistemic uncertainty.

4.4. Evaluation Protocol

The performance of the models will be evaluated using a comprehensive suite of metrics to assess both classification accuracy and uncertainty quality.Evaluation will be performed on the held-out test set of ChestX-ray14 to assess model generalization. For benchmarking against other studies, performance may also be compared on CheXpert and MIMIC-CXR datasets.

- **Classification Metrics:**
  - **AUROC (Area Under the Receiver Operating Characteristic Curve):** The primary metric for multi-label classification on imbalanced datasets. Per-class AUROC scores will be calculated to understand performance on each of the 14 diseases individually, in addition to micro and macro-averaged scores for a holistic view [30].
  - **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of the model's performance on both positive and negative predictions, particularly important for rare conditions. Per-class F1 scores will also be calculated to assess performance on each disease.
- **Uncertainty Metrics:**
  - **Expected Calibration Error (ECE):** This metric measures how well the model's predicted probabilities align with its true accuracy, with a lower ECE indicating a more reliable and well calibrated model [31].
  - **Negative Log-Likelihood (NLL):** A standard loss function that doubles as an evaluation metric, penalizing incorrect and overconfident predictions. A lower NLL indicates a more trustworthy model [32].

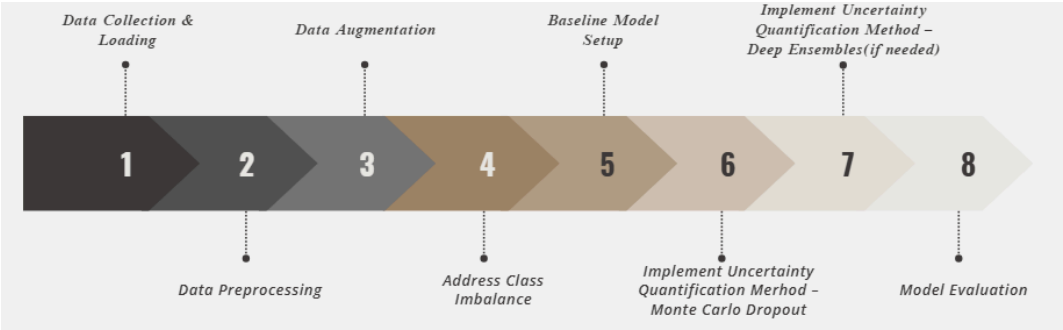

Figure 3: Workflow of the Proposed Methodology

5. Project Timeline

Phase 1: Literature Review & Methodology Outline (Weeks 5–6)

- **Tasks:**
  - Conduct a comprehensive literature review and critically evaluate existing models.
  - Prepare a detailed methodology outline to guide subsequent phases.

Phase 2: Baseline Replication (Week 7)

- **Tasks:**
  - Replicate CheXNet (DenseNet-121) on the benchmark dataset.
  - Establish reproducible baseline performance using AUROC and F1-score.

Phase 3: Methodology Implementation (Weeks 8–9)

- **Tasks:**
  - Implement data augmentation and class imbalance mitigation strategies.
  - Integrate Monte Carlo Dropout into DenseNet-121.
  - Develop and train Deep Ensemble of CheXNet models (if needed).
  - Prepare a short paper summarizing preliminary results.

Phase 4: Rigorous Evaluation & Analysis (Weeks 10–11)

- **Tasks:**
  - Evaluate models using AUROC, F1-score, Expected Calibration Error (ECE), and Negative Log-Likelihood (NLL).
  - Compare Monte Carlo Dropout and/or Deep Ensembles (if implemented) in terms of uncertainty estimation and calibration.

Phase 5: Final Report & Conference Submission (Week 12)

- **Tasks:**
  - Prepare the final research paper in conference format, including the abstract, introduction, methodology, results, and conclusions.
  - Submit the paper for conference review.

6. Conclusion

This project aims to bridge a critical gap in clinical AI by enhancing the CheXNet model with robust uncertainty quantification capabilities. Through the implementation of Monte Carlo Dropout and, if necessary, Deep Ensembles, my objective is to transform the model from a basic prediction tool into a reliable clinical decision support tool. This work will demonstrate that the true value of an AI model in a high-stakes environment is not solely based on its raw classification metrics, but on its ability to

communicate its own confidence. A model that can reliably flag an ambiguous case for an expert clinician is, arguably, more valuable than one that is slightly more accurate on average but provides no measure of confidence.

References

[1] L. Huang, S. Ruan, Y. Xing, and M. Feng, "A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods," *Med. Image Anal.*, vol. 97, p. 103223, Oct. 2024, doi: 10.1016/j.media.2024.103223.

[2] ~ Laurenoakdenrayner, "CheXNet: an in-depth review," Lauren Oakden-Rayner. Accessed: Aug. 24, 2025. [Online]. Available: https://laurenoakdenrayner.com/2018/01/24/chexnet-an-in-depth-review/

[3] "Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations".

[4] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Dec. 25, 2017, *arXiv*: arXiv:1711.05225. doi: 10.48550/arXiv.1711.05225.

[5] "NIH Chest X-rays." Accessed: Aug. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/nih-chest-xrays/data

[6] S. Iqbal, A. N. Qureshi, J. Li, I. A. Choudhry, and T. Mahmood, "Dynamic learning for imbalanced data in learning chest X-ray and CT images," *Heliyon*, vol. 9, no. 6, p. e16807, June 2023, doi: 10.1016/j.heliyon.2023.e16807.

[7] "Stanford AIMI Shared Datasets." Accessed: Aug. 24, 2025. [Online]. Available: https://stanfordaimi.azurewebsites.net/datasets/5158c524-d3ab-4e02-96e9-6ee9efc110a1

[8] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, "Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks," Apr. 24, 2018, *arXiv*: arXiv:1804.07839. doi: 10.48550/arXiv.1804.07839.

[9] M. I. U. Haque, A. K. Dubey, I. Danciu, A. C. Justice, O. S. Ovchinnikova, and J. D. Hinkle, "Effect of image resolution on automated classification of chest X-rays," *J. Med. Imaging*, vol. 10, no. 4, p. 044503, July 2023, doi: 10.1117/1.JMI.10.4.044503.

[10] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 24, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[11] S. Baur, W. Samek, and J. Ma, "Benchmarking Uncertainty and its Disentanglement in multi-label Chest X-Ray Classification," Aug. 06, 2025, *arXiv*: arXiv:2508.04457. doi: 10.48550/arXiv.2508.04457.

[12] M. A. Chan, M. J. Molina, and C. A. Metzler, "Estimating Epistemic and Aleatoric Uncertainty with a Single Model".

[13] L. Ngartera, M. A. Issaka, and S. Nadarajah, "Application of Bayesian Neural Networks in Healthcare: Three Case Studies," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 4, pp. 2639–2658, Dec. 2024, doi: 10.3390/make6040127.

[14] C. C. Margossian, L. Pillaud-Vivien, and L. K. Saul, "Variational Inference for Uncertainty Quantification: an Analysis of Trade-offs," May 06, 2025, *arXiv*: arXiv:2403.13748. doi: 10.48550/arXiv.2403.13748.

[15] M. Hasan, A. Khosravi, I. Hossain, A. Rahman, and S. Nahavandi, "Controlled Dropout for Uncertainty Estimation," May 06, 2022, *arXiv*: arXiv:2205.03109. doi: 10.48550/arXiv.2205.03109.

[16] "Deep ensembles - AWS Prescriptive Guidance." Accessed: Aug. 24, 2025. [Online]. Available: https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/deep-ensembles.html
[17] A. Whata, K. Dibeco, K. Madzima, and I. Obagbuwa, "Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia," *Front. Artif. Intell.*, vol. 7, Sept. 2024, doi: 10.3389/frai.2024.1410841.
[18] "Deep ensembles - AWS Prescriptive Guidance." Accessed: Aug. 24, 2025. [Online]. Available: https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/deep-ensembles.html
[19] S. Lee, "Advanced Uncertainty Estimation Methods." Accessed: Aug. 24, 2025. [Online]. Available: https://www.numberanalytics.com/blog/advanced-uncertainty-estimation-methods-medical-imaging
[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 24, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html
[21] A. Whata, K. Dibeco, K. Madzima, and I. Obagbuwa, "Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia," *Front. Artif. Intell.*, vol. 7, p. 1410841, Sept. 2024, doi: 10.3389/frai.2024.1410841.
[22] "Estimating Epistemic and Aleatoric Uncertainty with a Single Model." Accessed: Aug. 24, 2025. [Online]. Available: https://arxiv.org/html/2402.03478v2
[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Aug. 03, 2017, *arXiv*: arXiv:1706.04599. doi: 10.48550/arXiv.1706.04599.
[24] X. Zeng, H. Wang, L. Zhao, Y. Cheng, D. Zhou, and S. Shi, "Uncertainty Quantification and Temperature Scaling Calibration for Protein-RNA Binding Site Prediction," *J. Chem. Inf. Model.*, vol. 65, no. 12, pp. 6310–6321, June 2025, doi: 10.1021/acs.jcim.5c00556.
[25] "(PDF) NIH-Chest-X-rays-Multi-Label-Image-Classification." Accessed: Aug. 24, 2025. [Online]. Available: https://www.researchgate.net/publication/391056217_NIH-Chest-X-rays-Multi-Label-Image-Classification
[26] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Dec. 25, 2017, *arXiv*: arXiv:1711.05225. doi: 10.48550/arXiv.1711.05225.
[27] D. Strick, C. Garcia, and A. Huang, "Reproducing and Improving CheXNet: Deep Learning for Chest X-ray Disease Classification," May 10, 2025, *arXiv*: arXiv:2505.06646. doi: 10.48550/arXiv.2505.06646.
[28] S. Rajaraman, G. Zamzmi, and S. K. Antani, "Novel loss functions for ensemble-based medical image classification," *PLOS ONE*, vol. 16, no. 12, p. e0261307, Dec. 2021, doi: 10.1371/journal.pone.0261307.
[29] "Advanced Uncertainty Estimation Methods." Accessed: Aug. 24, 2025. [Online]. Available: https://www.numberanalytics.com/blog/advanced-uncertainty-estimation-methods-medical-imaging
[30] S. Baur, W. Samek, and J. Ma, "Benchmarking Uncertainty and its Disentanglement in multi-label Chest X-Ray Classification," Aug. 06, 2025, *arXiv*: arXiv:2508.04457. doi: 10.48550/arXiv.2508.04457.
[31] M. Barandas *et al.*, "Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram," *Inf. Fusion*, vol. 101, p. 101978, Jan. 2024, doi: 10.1016/j.inffus.2023.101978.
[32] L. Nieradzik, "Metrics for uncertainty estimation," lars76.github.io. Accessed: Aug. 24, 2025. [Online]. Available: https://lars76.github.io/2020/08/07/metrics-for-uncertainty-estimation.html