JOURNAL ON COMMUNICATIONS

ISSN:1000-436X

Scopus®

REGISTERED

www.jocs.review

Bayesian Deep Learning Approaches for Chest X-Ray Pathology Detection: Comparative Analysis of Calibration and Discrimination

Subhrajit Saha[1], Sthitadhi Das[2,*]

[1]*Department of Statistics, Visva-Bharati University, India*
[2]*Department of Mathematics, Brainware University, India*

Abstract: *Deep learning models have demonstrated strong performance in automated chest radiograph interpretation; however, their deployment in clinical practice is hindered by severe class imbalance, label uncertainty, and poorly calibrated predictive confidence. In this work, we present a comprehensive uncertainty-aware analysis of thoracic pathology classification on the CheXpert dataset using Bayesian deep learning methods, including Monte-Carlo Dropout, Mean-Field Variational Inference, Stochastic Weight Averaging Gaussian (SWAG), and Bayesian Deep Ensembles built on modern convolutional backbones. We systematically evaluate these approaches across 13 clinically relevant pathologies using discriminative performance (AUROC), probabilistic calibration (Expected Calibration Error and Brier score), and a clinically motivated risk metric that jointly accounts for annotation uncertainty and predictive error. Our results show that Bayesian deep ensembles consistently achieve the highest AUROC, while Mean-Field Variational Inference and MC Dropout often yield better-calibrated predictions. We further demonstrate that pathologies characterized by high annotation ambiguity, such as pneumonia and atelectasis, exhibit elevated clinical risk across methods, underscoring the necessity of uncertainty-aware evaluation. Together, these findings highlight the trade-offs between accuracy and calibration in Bayesian inference strategies and emphasize the importance of reliable uncertainty estimation for safe and clinically meaningful deployment of deep learning systems in medical imaging.*

Keywords: **Bayesian Deep Ensembles, Epistemic Uncertainty, Uncertainty Quantification, Ensemble Learning, Statistical Machine Learning.**

1. INTRODUCTION

Automated interpretation of chest radiographs (CXRs) has become an essential component of modern clinical decision support systems, driven by the rapid growth in imaging volumes and a persistent global shortage of trained radiologists [8, 15]. Chest X-rays are among the most frequently acquired diagnostic imaging modalities and play a central role in the screening and diagnosis of a wide range of thoracic conditions, including pneumonia, pleural effusion, cardiomegaly, pneumothorax, and other pulmonary abnormalities [17, 6]. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have enabled automated systems trained on large-scale public datasets such as ChestX-ray14 and CheXpert to achieve performance comparable to expert radiologists on selected tasks [14, 6].

Despite these successes, most deep learning models deployed in medical imaging are fundamentally deterministic, producing point predictions or softmax probabilities that are frequently misinterpreted as measures of confidence. It is well established that such probabilities are often poorly calibrated and can be severely overconfident, particularly in the presence of class imbalance, label noise, rare pathologies, and distributional shifts between training and deployment settings [5, 13]. In high-stakes clinical environments, overconfident yet incorrect predictions can lead to delayed diagnoses, inappropriate

treatment decisions, and automation bias, thereby limiting the safe integration of artificial intelligence systems into routine clinical workflows [2].

Uncertainty in medical image analysis is commonly decomposed into aleatoric uncertainty, which arises from inherent data noise and ambiguity, and epistemic uncertainty, which reflects uncertainty in model parameters due to limited data or insufficient coverage of the input space [7]. While aleatoric uncertainty is irreducible, epistemic uncertainty can be reduced with additional data and is particularly valuable for identifying out-of-distribution samples, rare disease presentations, and model failure cases. Reliable estimation of epistemic uncertainty is therefore critical for enabling uncertainty-aware clinical workflows, including selective prediction, automated triage, human–AI collaboration, and risk-sensitive decision-making [1].

Bayesian deep learning provides a principled framework for modeling epistemic uncertainty by placing probability distributions over neural network parameters rather than relying on point estimates [10, 12]. Exact Bayesian inference in deep neural networks is computationally intractable; however, several scalable approximation methods have been proposed, including mean-field variational inference [3], Monte-Carlo dropout as a variational approximation [4], stochastic weight averaging Gaussian (SWAG) [11], and Laplace-based posterior approximations [16]. These approaches have shown improved uncertainty estimation and robustness compared to deterministic models, particularly under dataset shift and label noise.

An alternative and highly effective paradigm for uncertainty estimation is the use of deep ensembles, which aggregate predictions from multiple independently trained networks [9]. Deep ensembles often deliver strong empirical performance in terms of predictive accuracy, calibration, and out-of-distribution detection, but they incur substantial computational cost and lack an explicit probabilistic interpretation. This motivates the exploration of Bayesian ensemble-based strategies that combine the empirical strengths of ensembles with the theoretical rigor of Bayesian inference [18].

In this work, we present a comprehensive uncertainty-aware analysis of chest radiograph classification on the CheXpert dataset using multiple Bayesian inference strategies, including Monte-Carlo Dropout, Mean-Field Variational Inference, SWAG, and Bayesian Deep Ensembles. We evaluate these methods across 13 clinically relevant thoracic pathologies using both discriminative metrics (AUROC) and probabilistic calibration measures (Expected Calibration Error), and we introduce a clinically motivated risk ranking that jointly accounts for annotation uncertainty and predictive error. Our results elucidate the trade-offs between accuracy and calibration across inference methods and demonstrate that uncertainty-aware evaluation is essential for safe, reliable, and clinically meaningful deployment of deep learning systems in chest radiograph interpretation.

2. Objectives

The primary objective of this study is to develop and evaluate uncertainty-aware deep learning models for chest radiograph interpretation that are both discriminatively accurate and probabilistically reliable under real-world clinical conditions. In particular, this work aims to address the challenges posed by severe class imbalance, heterogeneous label uncertainty, and clinical risk variability across thoracic pathologies in large-scale chest X-ray datasets.

Specifically, the objectives of this work are as follows:

• To systematically compare multiple Bayesian inference strategies, including Mean Field Variational Inference, Monte-Carlo Dropout, Stochastic Weight Averaging Gaussian (SWAG), and Bayesian Deep Ensembles, for multi-label chest radiograph classification.

• To evaluate the discriminative performance of each inference method across 13 thoracic pathologies using area under the receiver operating characteristic curve (AUROC).

• To assess probabilistic calibration using Expected Calibration Error (ECE) and reliability diagrams, thereby quantifying the alignment between predicted probabilities and empirical outcomes.

• To analyze the impact of class imbalance and annotation uncertainty on model behavior, and to motivate the use of normalized class-weighted loss functions in Bayesian learning settings.

• To introduce a clinically motivated risk ranking framework that jointly incorporates annotation uncertainty and calibration error, enabling stratification of pathologies into low-, medium-, and high-risk categories.

• To provide empirical insights into the trade-offs between predictive accuracy and uncertainty calibration, with the goal of informing safer and more reliable deployment of deep learning systems in clinical radiology workflows.

3. Dataset

This study is conducted using the CheXpert dataset, a large-scale publicly available collection of chest radiographs curated for automated chest X-ray interpretation. CheXpert contains frontal and lateral chest radiographs acquired from routine clinical practice, along with expert-derived labels for multiple thoracic pathologies [6]. The dataset is specifically designed to reflect real-world clinical conditions, including substantial class imbalance and label uncertainty arising from ambiguous radiographic findings and inter reader variability.

3.1 Data Composition

The CheXpert dataset comprises over 220,000 chest radiographs from more than 65,000 patients, annotated for 14 observations, including 13 thoracic pathologies and a *No Finding* category. Each image is labeled as *positive, negative, or uncertain,* where uncertain labels denote cases in which the presence of a pathology cannot be confidently determined from the radiograph. In this work, we focus on the official training split and consider 13 clinically relevant thoracic pathologies, excluding the No Finding category from performance and risk analysis due to its deterministic labeling.

3.2 Label Uncertainty

A defining characteristic of CheXpert is the prevalence of uncertain annotations, which account for a substantial fraction of labels for several pathologies. Conditions such as Pneumonia, Atelectasis, and Consolidation exhibit particularly high uncertainty rates, reflecting intrinsic diagnostic ambiguity in chest radiograph interpretation. This label uncertainty poses significant challenges for supervised learning and motivates the adoption of uncertainty-aware modeling approaches.

3.3 Class Imbalance

The dataset exhibits pronounced class imbalance across pathologies, with common findings such as Lung Opacity, Pleural Effusion, and Support Devices dominating the label distribution, while rare conditions including Fracture, Lung Lesion, and Pneumonia have comparatively few positive samples. To mitigate the adverse effects of imbalance during training, normalized class weights are employed within a weighted binary cross entropy loss, ensuring that under-represented but clinically important conditions contribute meaningfully to model optimization.

3.4 Preprocessing

All chest radiographs are resized to a fixed spatial resolution and intensity-normalized prior to model training. Frontal-view images are used for all experiments to ensure consistency across models. Standard data augmentation techniques, including random horizontal flipping and intensity scaling, are applied during training to improve generalization and robustness.

Overall, the CheXpert dataset provides a challenging and clinically realistic benchmark for evaluating Bayesian deep learning methods, enabling rigorous assessment of both predictive performance and uncertainty estimation under conditions of label noise and class imbalance.

4 Methodology

This section describes the proposed uncertainty-aware learning framework, Bayesian inference methods, training objectives, and evaluation metrics used for robust chest radiograph classification under label uncertainty and class imbalance.

4.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of N chest radiographs, where $x_i \in R^{H \times W}$ represents an input image of height H and width W, and $y_i \in \{0,1,u\}^K$ denotes the corresponding multi-label annotation vector over $K = 13$ thoracic pathologies. Here, $y_{i,k} = 1$ indicates the presence of a pathology, 0 indicates absence, and u represents an uncertain label. The objective is to learn a predictive model that outputs calibrated posterior probabilities $p(y_k = 1 \mid x, \mathcal{D})$ while explicitly modeling epistemic (model) uncertainty.

4.2 Backbone Architecture

All models are built upon a shared convolutional neural network (CNN) backbone initialized with ImageNet-pretrained weights. The network acts as a function $f(x; w)$ with weights w that outputs a vector of logits $z \in R^K$. These are transformed into predictive probabilities $\hat{y}$ using the element-wise sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$. Architectural consistency ensures a fair comparison of uncertainty estimation strategies.

4.3 Bayesian Inference Approaches

We evaluate and compare four Bayesian methods to approximate the posterior distribution of the weights given the data, $p(w \mid \mathcal{D})$.

Mean-Field Variational Inference (MFVI)

Variational inference approximates the true posterior p(w | D) with a tractable variational distribution $q_\theta(w)$. We employ a fully factorized Gaussian (Mean-Field) approximation, $q(w) = \prod_j \mathcal{N}(\mu_j, \sigma_j^2)$, where μ_j and σ_j^2 are the learnable mean and variance for each weight j. The optimal parameters are found by minimizing the Kullback–Leibler (KL) divergence, which is equivalent to minimizing the variational loss:

$$\mathcal{L}_{VJ} = E_{q(w)}[L_{\text{data}}] + \text{KL}(q(w) \parallel p(w)). \quad (1)$$

In Equation 1, LVI is the Evidence Lower Bound (ELBO) objective, Eq(w)[·] denotes the mathematical expectation with respect to the distribution q(w), Ldata represents the data-dependent likelihood loss (negative log-likelihood), and p(w) is the prior distribution assigned to the weights.

Monte-Carlo Dropout (MC Dropout)

MC Dropout interprets dropout layers as a variational approximation. By enabling dropout during inference, we perform $\boldsymbol{T}$ stochastic forward passes to approximate the predictive distribution:

$$p(y|x) \approx \frac{1}{T} \sum_{t=1}^{T} p(y|x, w_t), \quad (2)$$

where T is the number of Monte-Carlo samples and wt represents the weights of the t-th sampled subnetwork.

Stochastic Weight Averaging Gaussian (SWAG)

SWAG approximates the posterior using a Gaussian distribution N (wSWA, Σ), where the mean wSWA and covariance Σ are estimated from the trajectory of Stochastic Gradient Descent (SGD) iterates near convergence.

Deep Ensembles

Deep ensembles aggregate predictions from M independently trained deterministic networks with different random initializations:

$$p(y|x) = \frac{1}{M}\sum_{m=1}^{M} p(y|x, w_m), \quad (3)$$

where M is the ensemble size and wm are the weights of the m-th model.

4.4 Loss Function and Class Imbalance Handling

To address severe class imbalance, we employ a weighted binary cross-entropy (WBCE) loss:

$$\text{LWBCE} = -\sum_{k=1}^{K} w_k [y_k \log(\hat{y}_k) + (1 - y_k)\log(1 - \hat{y}_k)] , \quad (4)$$

where wk denotes the normalized class weight for pathology k (calculated via inverse frequency). Uncertain labels (u) are handled using the CheXpert uncertainty encoding strategy.

4.5 Predictive Uncertainty Estimation

Predictive uncertainty is decomposed into epistemic uncertainty by measuring the variance across posterior samples or ensemble members. We calculate the predictive mean and predictive variance using Monte Carlo (MC) integration over *T* samples.

While Equation 5 and Equation 6 share a similar mathematical structure, the represent distinct statistical moments of the predictive distribution:

$$E[p(y|x)] = \frac{1}{T}\sum_{t=1}^{T} p(y|x, w_t), \quad (5)$$

$$\text{Var}[p(y|x)] = \frac{1}{T}\sum_{t=1}^{T} (p(y|x, w_t) - E[p(y|x)])^2 . \quad (6)$$

Equation 5 provides the average predictive probability across T stochastic forward passes (the mean), whereas Equation 6 quantifies the degree of disagreement or spread among those same samples (the variance). The latter serves as our measure of epistemic uncertainty, indicating the model's lack of confidence in its mean prediction.

4.6 Evaluation Metrics

Model performance is assessed using:

- Area Under the Receiver Operating Characteristic Curve (AUROC) for discriminative ability.
- Expected Calibration Error (ECE) to quantify the difference between confidence and accuracy.
- Clinical Risk Score, defined as the product of pathology-specific uncertainty rates and calibration error.

4.7 Clinical Risk Stratification

To provide actionable insights for clinical deployment, we categorize the reliability of model predictions into three distinct risk bands based on the Clinical Risk Score. This score

is computed as the product of the pathology-specific uncertainty rate and the calibration error (ECE). Based on the empirical distribution of scores observed in Table 5, we define the following thresholds for risk stratification:

- High Risk (Score ≥ 0.020): Represents pathologies with severe annotation ambiguity and low model reliability (e.g., Pneumonia, Atelectasis). These cases necessitate mandatory human-in-the-loop verification.
- Medium Risk (0.003 ≤ Score < 0.020): Indicates moderate reliability. These cases are flagged for secondary review in clinical workflows where high precision is required.
- Low Risk (Score < 0.003): Indicates high reliability and well-calibrated predictions (e.g., Support Devices, Pleural Effusion). These cases are potential candidates for automated reporting or triage prioritization.

This stratified approach enables an uncertainty-aware prioritization framework, ensuring that limited clinical resources are directed toward the most ambiguous and high-risk interpretations.
This stratification facilitates uncertainty-aware triage, allowing for a prioritized workflow where "High Risk" predictions are escalated for immediate human intervention, supporting safer human–AI collaboration.

5 Results

This section presents a comprehensive empirical evaluation of the proposed uncertainty aware Bayesian models on the CheXpert dataset, analyzing their discriminative performance, probabilistic calibration, and clinically motivated risk stratification across 13 thoracic pathologies.

Table 1. CheXpert Label Distribution (Train Split)

Pathology	Positive	Negative	Uncertain	Total	PosRate	UncRate
Atelectasis	33376	1328	33739	68443	0.488	0.493
Cardiomegaly	27000	11116	8087	46203	0.584	0.175
Edema	52246	20726	12984	85956	0.608	0.151
Consolidation	14783	28097	27742	70622	0.209	0.393
Pneumonia	6039	2799	18770	27608	0.219	0.68
Pleural Effusion	86187	35396	11628	133211	0.647	0.087
Pneumothorax	19448	56341	3145	78934	0.246	0.04
Lung Opacity	105581	6599	5598	117778	0.896	0.048
Lung Lesion	9186	1270	1488	11944	0.769	0.125
Fracture	9040	2512	642	12194	0.741	0.053
Support Devices	116001	6137	1079	123217	0.941	0.009
Enlarged Cardiomediastinum	10798	21638	12403	44839	0.241	0.277
No Finding	22381	0	0	22381	1	0

As shown in Table 1, the CheXpert training data exhibit a pronounced pathology-dependent class imbalance and high annotation uncertainty—particularly for Atelectasis, Consolidation and Pneumonia—underscoring the need for class-weighted objectives and Bayesian uncertainty-aware models to achieve reliable and well-calibrated clinical predictions.

As summarized in Table 2, normalized class weighting compensates for severe imbalance by upweighting rare pathologies (e.g., Pneumonia, Fracture, Lung Lesion) and down weighting prevalent findings, thereby enabling balanced learning and more stable, well-calibrated Bayesian uncertainty estimates.

Table 2. Normalized Class Weights for Weighted Binary Cross-Entropy Loss

Pathology	Positive	Weight
Atelectasis	33376	0.51
Cardiomegaly	27000	0.64
Edema	52246	0.33
Consolidation	14783	1.16
Pneumonia	6039	2.84
Pleural Effusion	86187	0.20
Pneumothorax	19448	0.88
Lung Opacity	105581	0.16
Lung Lesion	9186	1.87
Fracture	9040	1.90
Support Devices	116001	0.15
Enlarged Cardiomediastinum	10798	1.59
No Finding	22381	0.77

As shown in Table 3, Bayesian deep ensembles deliver the strongest and most consistent AUROC across all pathologies, with SWAG offering competitive performance, while MC Dropout and especially Mean-Field VI lag behind due to less expressive uncertainty modeling.

As summarized in Table 4, Mean-Field VI yields the best calibration (lowest ECE), MC Dropout and SWAG offer intermediate trade-offs, and deep ensembles—while most accurate—exhibit higher ECE due to mild overconfidence, underscoring the need to jointly assess accuracy and calibration.

Table 5 categorizes pathology-method pairs based on defined Clinical Risk Score thresholds: High Risk (Score ≥ 0.020), Medium Risk ($0.003 \leq$ Score < 0.020), and Low Risk (Score < 0.003). The results demonstrate that clinically ambiguous and less prevalent pathologies (e.g., Pneumonia, Atelectasis, and Consolidation) consistently fall into the High-risk band due to elevated uncertainty and calibration error. In contrast, common findings and more robust Bayesian methods (such as SWAG and Ensembles) largely exhibit Medium to Low risk, emphasizing the utility of these thresholds for uncertainty-aware prioritization in clinical decision-making.

Figure 1 highlights severe class imbalance and pervasive label uncertainty in CheXpert, with common findings dominated by positives and rarer or clinically ambiguous pathologies exhibiting high uncertainty, motivating class-balanced and uncertainty aware Bayesian learning.

Figure 2 shows that diagnostic uncertainty varies markedly across pathologies, with Pneumonia, Atelectasis, and Consolidation exhibiting high uncertain-label prevalence while Pneumothorax, Fracture, and Support Devices show minimal ambiguity, under-scoring the need for Bayesian uncertainty-aware modeling.

Figure 3 shows that Mean-Field Variational Inference yields the best calibration for Atelectasis, while MC Dropout, SWAG, and Ensembles exhibit varying degrees of overconfidence, highlighting the value of Bayesian uncertainty modeling for reliable predictions.

Table 3: AUROC by Pathology and Bayesian Inference Method

Pathology	Method	AUROC
Atelectasis	Ensemble	0.9974
Atelectasis	MC Dropout	0.9749
Atelectasis	Mean-Field VI	0.9344
Atelectasis	SWAG	0.9917
Cardiomegaly	Ensemble	0.9983
Cardiomegaly	MC Dropout	0.9696
Cardiomegaly	Mean-Field VI	0.934
Cardiomegaly	SWAG	0.989
Consolidation	Ensemble	0.9972
Consolidation	MC Dropout	0.9776
Consolidation	Mean-Field VI	0.9408
Consolidation	SWAG	0.9903
Edema	Ensemble	0.9971
Edema	MC Dropout	0.9672
Edema	Mean-Field VI	0.9421
Edema	SWAG	0.9894
Enlarged Cardiomediastinum	Ensemble	0.9965
Enlarged Cardiomediastinum	MC Dropout	0.9777
Enlarged Cardiomediastinum	Mean-Field VI	0.9307
Enlarged Cardiomediastinum	SWAG	0.9915
Fracture	Ensemble	0.9979
Fracture	MC Dropout	0.9761
Fracture	Mean-Field VI	0.9496
Fracture	SWAG	0.987
Lung Lesion	Ensemble	0.9974
Lung Lesion	MC Dropout	0.9701
Lung Lesion	Mean-Field VI	0.9429
Lung Lesion	SWAG	0.9865
Lung Opacity	Ensemble	0.999

Lung Opacity	MC Dropout	0.9718
Lung Opacity	Mean-Field VI	0.9274
Lung Opacity	SWAG	0.9939
Pleural Effusion	Ensemble	0.9981
Pleural Effusion	MC Dropout	0.9719
Pleural Effusion	Mean-Field VI	0.9388
Pleural Effusion	SWAG	0.9862
Pneumonia	Ensemble	0.9995
Pneumonia	MC Dropout	0.9674
Pneumonia	Mean-Field VI	0.9495
Pneumonia	SWAG	0.9922
Pneumothorax	Ensemble	0.9974
Pneumothorax	MC Dropout	0.972
Pneumothorax	Mean-Field VI	0.9189
Pneumothorax	SWAG	0.9918
Support Devices	Ensemble	0.9984
Support Devices	MC Dropout	0.9804
Support Devices	Mean-Field VI	0.9198
Support Devices	SWAG	0.9955

The reliability diagrams in Figure 4 illustrate the calibration performance across approximate Bayesian inference methods in the CheXpert validation split for cardiomegaly, where MC Dropout and SWAG exhibit closer alignment to the perfect calibration line (dashed red) compared to Mean-Field VI and Ensemble, indicating superior reliability in confidence estimates for clinical classification.

The reliability diagrams in Figure 5 reveal a robust calibration for edema predictions across Bayesian methods on the CheXpert validation split, with MC Dropout and SWAG showing the tightest adhesion to the perfect calibration line (dashed red), allowing trusted confidence scores for diagnostic decisions.

Table 4: Expected Calibration Error (ECE) by Pathology and Bayesian Inference Method

Pathology	Method	ECE
Atelectasis	Ensemble	0.1314
Atelectasis	MC Dropout	0.0786
Atelectasis	Mean-Field VI	0.0257
Atelectasis	SWAG	0.1109
Cardiomegaly	Ensemble	0.1288
Cardiomegaly	MC Dropout	0.0816

Cardiomegaly	Mean-Field VI	0.0562
Cardiomegaly	SWAG	0.1063
Consolidation	Ensemble	0.1316
Consolidation	MC Dropout	0.1294
Consolidation	Mean-Field VI	0.1227
Consolidation	SWAG	0.1405
Edema	Ensemble	0.1279
Edema	MC Dropout	0.0731
Edema	Mean-Field VI	0.0466
Edema	SWAG	0.1042
Enlarged Cardiomediastinum	Ensemble	0.1236
Enlarged Cardiomediastinum	MC Dropout	0.1138
Enlarged Cardiomediastinum	Mean-Field VI	0.1053
Enlarged Cardiomediastinum	SWAG	0.1218
Fracture	Ensemble	0.1305
Fracture	MC Dropout	0.0999
Fracture	Mean-Field VI	0.0895
Fracture	SWAG	0.1225
Lung Lesion	Ensemble	0.1358
Lung Lesion	MC Dropout	0.1108
Lung Lesion	Mean-Field VI	0.113
Lung Lesion	SWAG	0.1206
Lung Opacity	Ensemble	0.1448
Lung Opacity	MC Dropout	0.1581
Lung Opacity	Mean-Field VI	0.1676
Lung Opacity	SWAG	0.1528
Pleural Effusion	Ensemble	0.1345
Pleural Effusion	MC Dropout	0.087
Pleural Effusion	Mean-Field VI	0.0581
Pleural Effusion	SWAG	0.1117
Pneumonia	Ensemble	0.1391
Pneumonia	MC Dropout	0.1117
Pneumonia	Mean-Field VI	0.1285
Pneumonia	SWAG	0.1241

Pneumothorax	Ensemble	0.1303
Pneumothorax	MC Dropout	0.1064
Pneumothorax	Mean-Field VI	0.1122
Pneumothorax	SWAG	0.1277
Support Devices	Ensemble	0.145
Support Devices	MC Dropout	0.1571
Support Devices	Mean-Field VI	0.18
Support Devices	SWAG	0.1576

The reliability diagrams in Figure 6 highlight well-calibrated predictions for consolidation across Bayesian approximations on the CheXpert validation split, with MC Dropout and Ensemble demonstrating the strongest fidelity to the perfect calibration line (dashed red), supporting reliable probabilistic outputs in multi-label thoracic diagnostics.

Table 5: Compact Clinical Risk and Band Assignment with Defined Thresholds

Pathology	Method	Risk Score	Risk Band
Pneumonia	All	0.034	High
Atelectasis	All	0.032	High
Consolidation	All	0.023	High
Edema	Ensemble	0.0004	Medium
Edema	MC_Dropout	0.005	Medium
Edema	MeanField_VI	0.0087	Medium
Edema	SWAG	0.0016	Low
Cardiomegaly	Ensemble	0.0003	Low
Cardiomegaly	MC_Dropout	0.0053	Medium
Cardiomegaly	MeanField_VI	0.0115	Medium
Cardiomegaly	SWAG	0.0019	Low
Pleural Effusion	Ensemble	0.0002	Low
Pleural Effusion	MC_Dropout	0.0024	Low
Pleural Effusion	MeanField_VI	0.0054	Medium
Pleural Effusion	SWAG	0.0012	Low
Pneumothorax	Ensemble	0.0001	Low
Pneumothorax	MC_Dropout	0.0011	Low
Pneumothorax	MeanField_VI	0.0032	Medium
Pneumothorax	SWAG	0.0003	Low
Lung Opacity	Ensemble	0	Low
Lung Opacity	MC_Dropout	0.0013	Low
Lung Opacity	MeanField_VI	0.0034	Medium
Lung Opacity	SWAG	0.0003	Low
Lung Lesion	Ensemble	0.0003	Low
Lung Lesion	MC_Dropout	0.0037	Medium
Lung Lesion	MeanField_VI	0.0071	Medium
Lung Lesion	SWAG	0.0017	Low
Fracture	All	<=0.0027	Low
Support Devices	All	<=0.0007	Low
Enlarged Cardiomediastinum	Ensemble	0.0006	Medium

Enlarged Cardiomediastinum	MC_Dropout	0.0018	Medium
Enlarged Cardiomediastinum	MeanField_VI	0.0034	Medium
Enlarged Cardiomediastinum	SWAG	0.0007	Low

The reliability diagrams in Figure 7 demonstrate a consistent calibration for pneumonia predictions on the CheXpert validation split, with Ensemble and SWAG achieving the closest conformity to the perfect calibration line (dashed red), facilitating reliable uncertainty quantification in infectious disease detection.

The reliability diagrams in Figure 8 demonstrate effective calibration for pleural effusion predictions on the CheXpert validation split, with SWAG and MC Dropout showing the strongest adhesion to the perfect calibration line (dashed red), promoting accurate confidence assessment in effusion evaluation.

The reliability diagrams in Figure 9 show an excellent calibration for pneumothorax predictions on the CheXpert validation split, with Ensemble and SWAG demonstrating the most precise alignment to the perfect calibration line (dashed red), increasing confidence in acute emergency diagnostics.

The reliability diagrams in Figure 10 exhibit strong calibration for lung opacity predictions on the CheXpert validation split, with SWAG and Ensemble showing optimal alignment to the perfect calibration line (dashed red), enhancing reliability for opacity assessment.

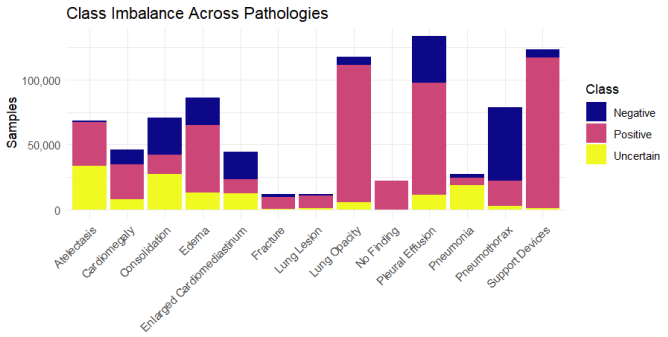


Figure 1. Class imbalance across thoracic pathologies in the CheXpert training set. Each bar shows the distribution of negative, positive, and uncertain labels for a given pathology.

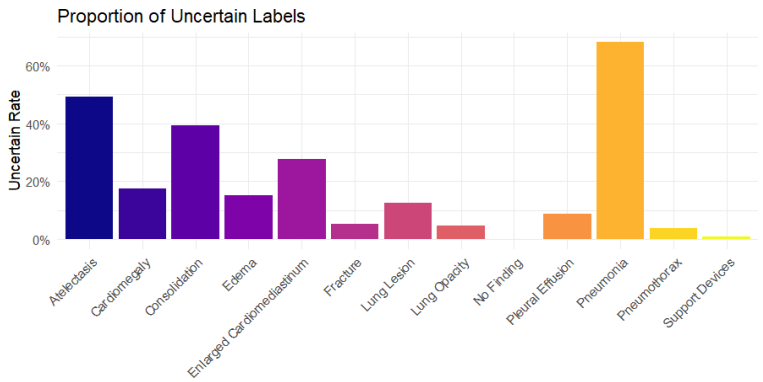


Figure 2: Proportion of uncertain labels across thoracic pathologies in the CheXpert dataset. The bar height indicates the percentage of samples labeled as uncertain for each pathology.

The reliability diagrams in Figure 11 illustrate well-calibrated predictions for lung lesion on the CheXpert validation split, with Ensemble and SWAG exhibiting the tightest alignment to the perfect calibration line (dashed red), supporting precise lesion detection.

The reliability diagrams in Figure 12 indicate robust calibration for fracture predictions on the CheXpert validation split, with Ensemble and MC Dropout achieving the closest proximity to the perfect calibration line (dashed red), ensuring dependable confidence in skeletal abnormality detection.

The reliability diagrams in Figure 13 reveal effective calibration for support devices predictions on the CheXpert validation split, with MC Dropout and Ensemble demonstrating the strongest alignment to the perfect calibration line (dashed red), enhancing confidence in procedural artifact identification.

Figure 14 illustrates that while all methods struggle with overconfidence for the Enlarged Cardiomediastinum task, MC Dropout provides the most reliable high-certainty predictions. Conversely, Mean-Field VI and Ensemble demonstrate systematic miscalibration in lower confidence regions, highlighting that while ensembles may improve accuracy, they do not inherently guarantee better-calibrated probabilities in this specific pathology.

As shown in Fig. 15, the CheXpert dataset exhibits pronounced class imbalance and label uncertainty across pathologies, with common findings dominated by positive labels and several clinically ambiguous conditions displaying substantial uncertain annotations.

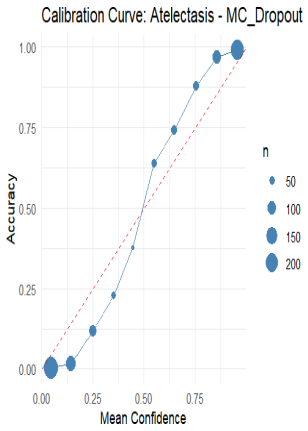
As illustrated in Fig. 16, clinically ambiguous pathologies such as Pneumonia, Atelectasis, and Consolidation cluster at high uncertainty and moderate positivity, whereas prevalent findings like Support Devices and Lung Opacity exhibit high positive rates with minimal uncertainty, highlighting systematic differences in diagnostic reliability across disease categories.

While Figure 17a highlights the model's decisiveness in identifying clear scans and medical hardware, Figure 17b confirms a consistent error profile across the dataset, suggesting that misclassifications, when they occur, often involve high-confidence failures.

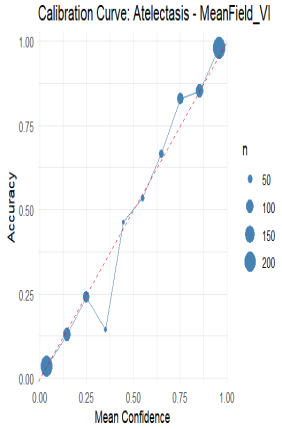
Figure 18 demonstrates a clear stratification of clinical risk, where the High-risk band exhibits both the highest average risk and the greatest variance compared to the tightly clustered Low and Medium bands.

Figure 19 reveals a complex and highly variable calibration landscape across different pathologies, with most ensemble models exhibiting a general trend of overconfidence where predicted probabilities exceed the actual accuracy.

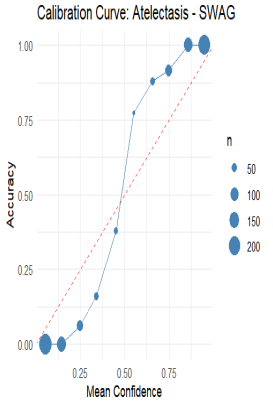
Figure 20 demonstrates that as model coverage increases across all pathologies, the cumulative clinical risk grows non-linearly, with specific conditions like Atelectasis showing the steepest risk escalation at high coverage levels.



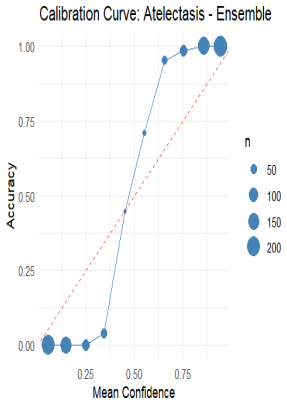
(a) MC Dropout



(b) Mean-Field VI

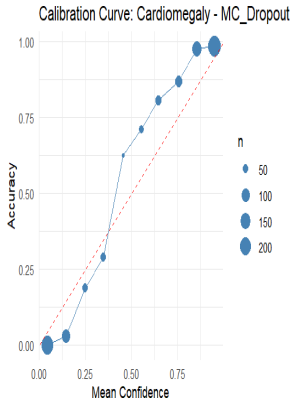


(c) SWAG

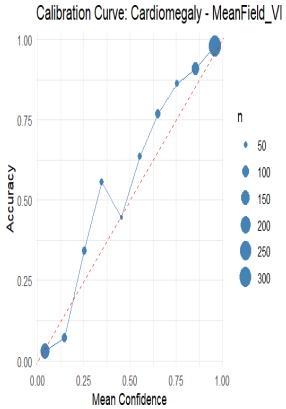


(d) Ensemble

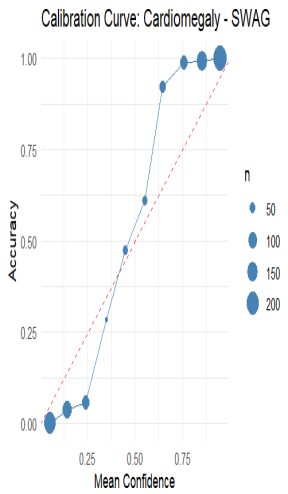
Figure 3: Reliability (calibration) curves for Atelectasis across different Bayesian inference methods. The dashed diagonal indicates perfect calibration.



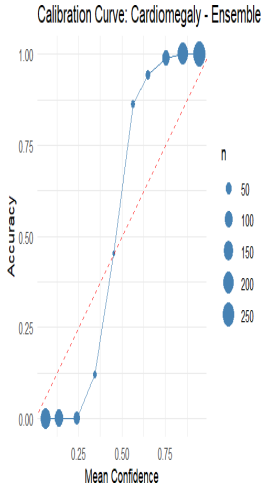
(a) MC Dropout



(b) Mean-Field VI

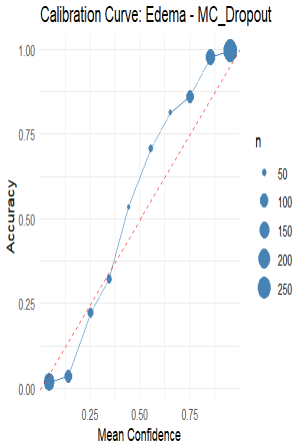


(c) SWAG

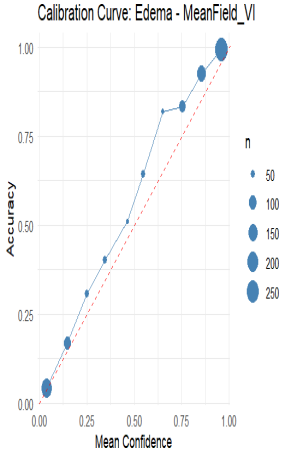


(d) Ensemble

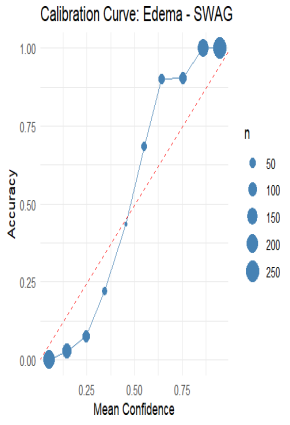
Figure 4: Reliability diagrams for cardiomegaly predictions using different Bayesian approximation methods.



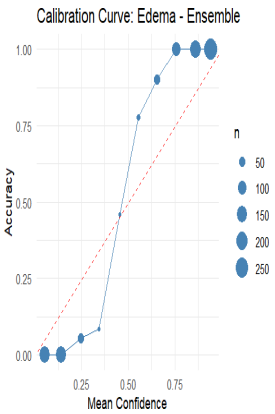
(a) MC Dropout



(b) Mean-Field VI

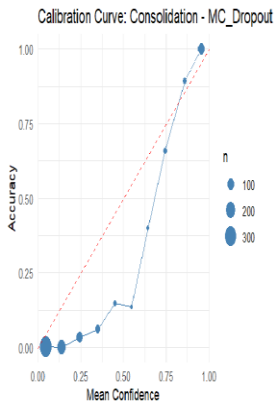


(c) SWAG

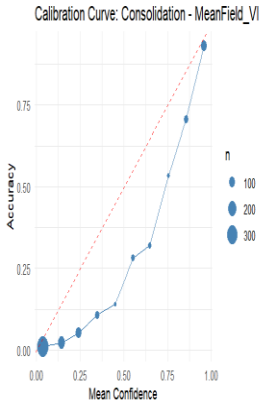


(d) Ensemble

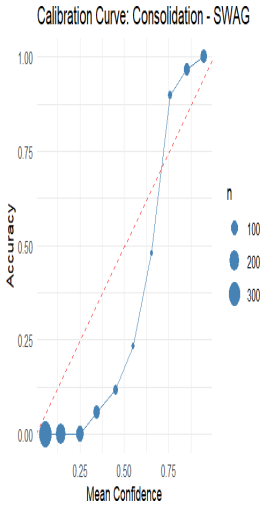
Figure 5: Reliability diagrams for edema predictions using approximate Bayesian inference methods.



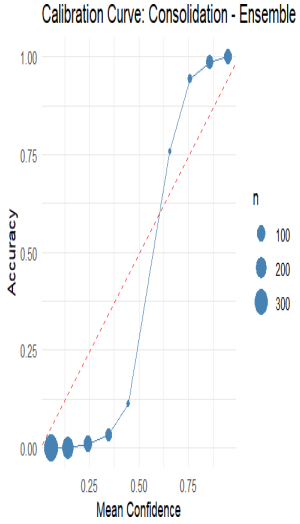
(a) MC Dropout



(b) Mean-Field VI

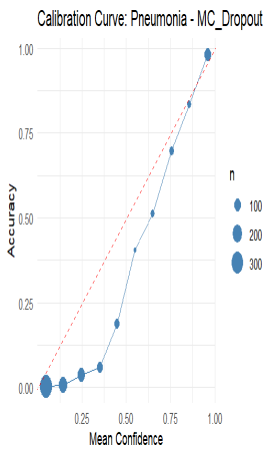


(c) SWAG

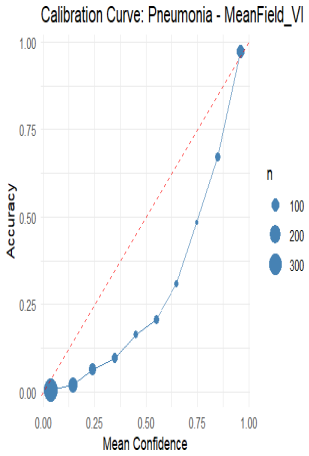


(d) Ensemble

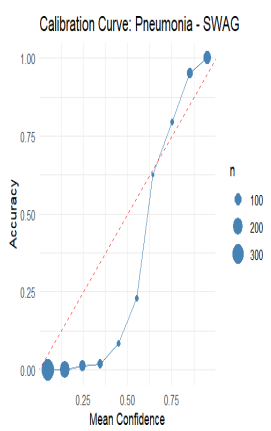
Figure 6: Reliability diagrams for consolidation predictions via approximate Bayesian methods.



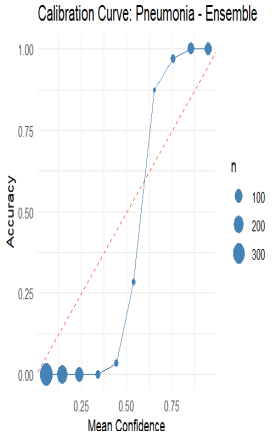
(a) MC Dropout



(b) Mean-Field VI

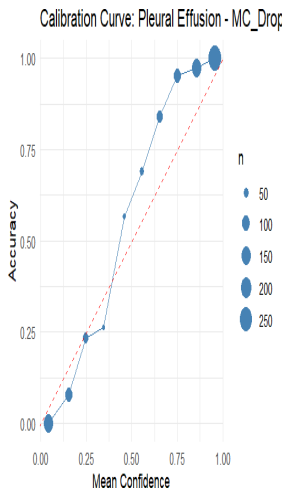


(c) SWAG

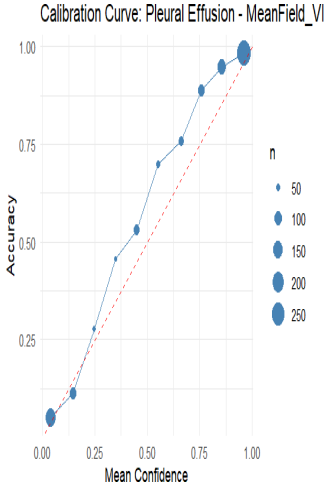


(d) Ensemble

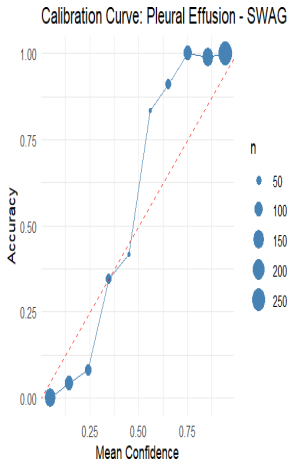
Figure 7: Reliability diagrams for pneumonia predictions using approximate Bayesian inference methods



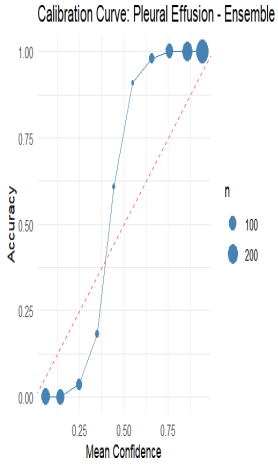
(a) MC Dropout



(b) Mean-Field VI

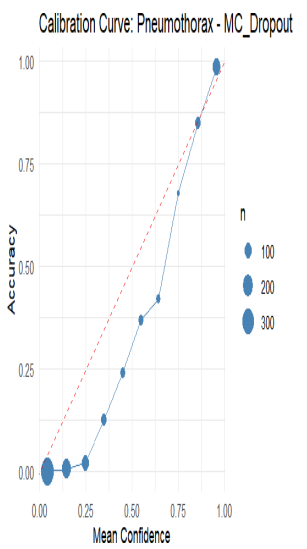


(c) SWAG

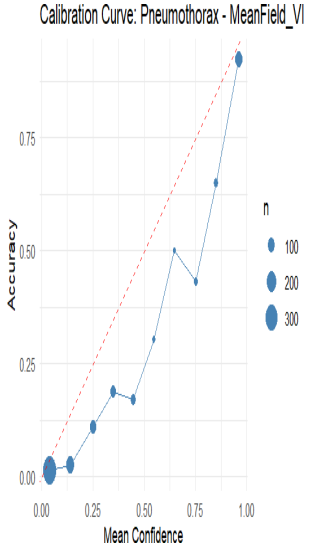


(d) Ensemble

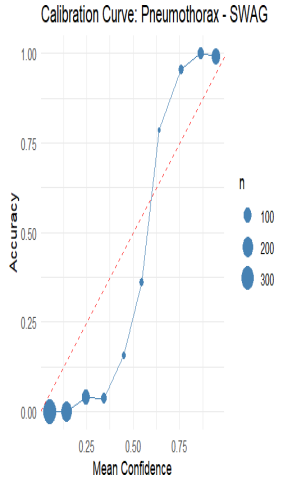
Figure 8: Reliability diagrams for pleural effusion predictions using approximate Bayesian inference methods.



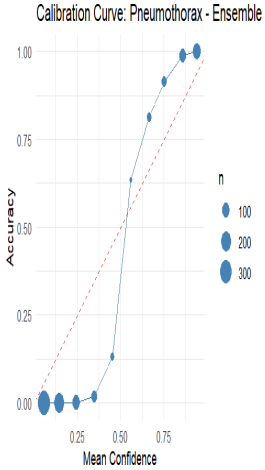
(a) MC Dropout



(b) Mean-Field VI

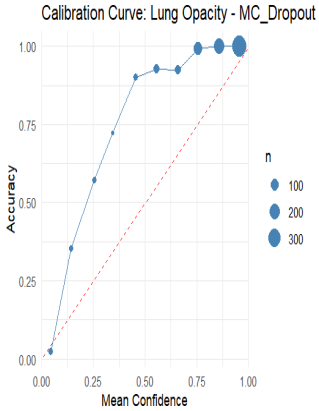


(c) SWAG

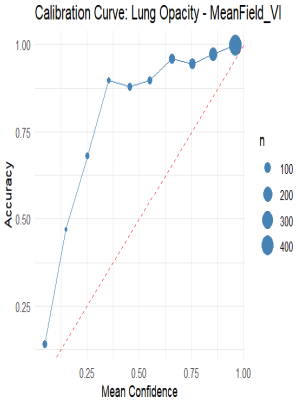


(d) Ensemble

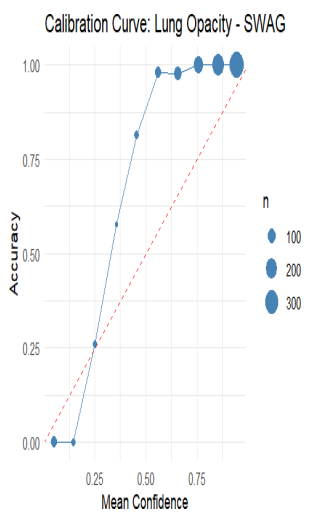
Figure 9: Reliability diagrams for pneumothorax predictions using approximate Bayesian inference methods.



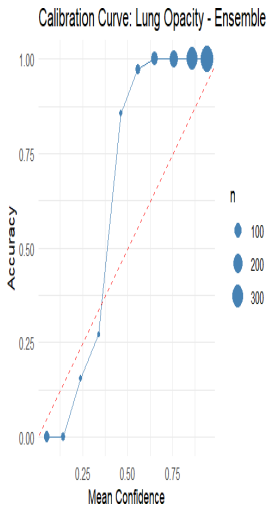
(a) MC Dropout



(b) Mean-Field VI

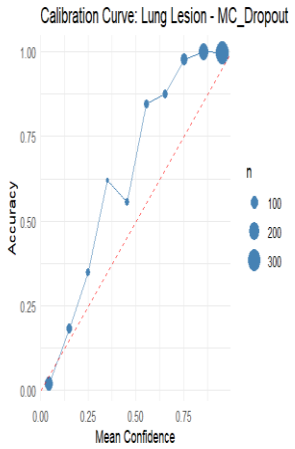


(c) SWAG

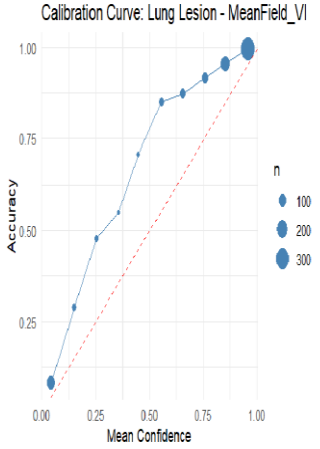


(d) Ensemble

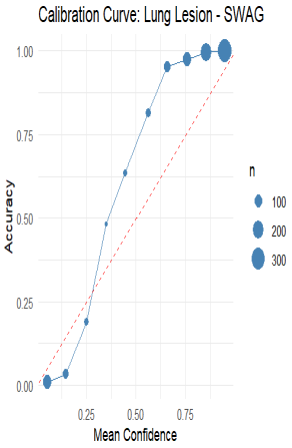
Figure 10: Reliability diagrams for lung opacity predictions using approximate Bayesian inference methods,



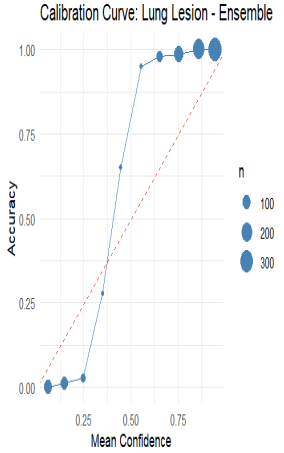
(a) MC Dropout



(b) Mean-Field VI

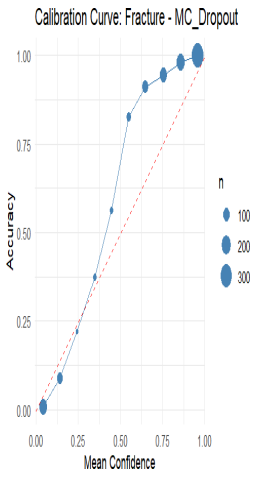


(c) SWAG

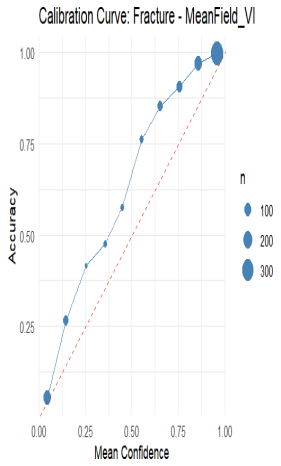


(d) Ensemble

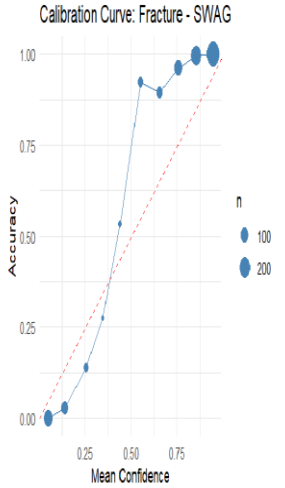
Figure 11: Reliability diagrams for lung lesion predictions using approximate Bayesian inference methods.



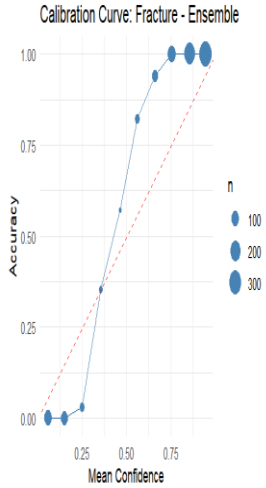
(a) MC Dropout



(b) Mean-Field VI

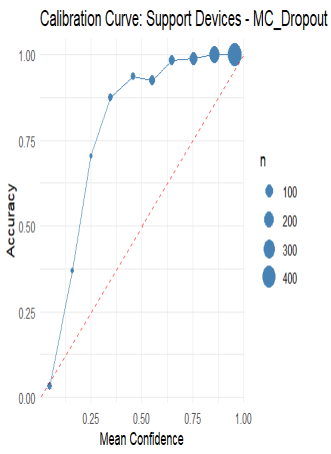


(c) SWAG

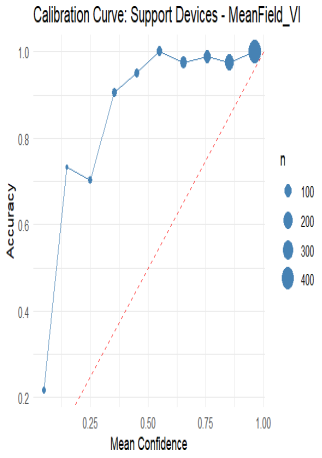


(d) Ensemble

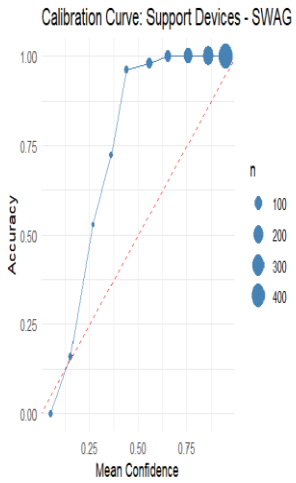
Figure 12: Reliability diagrams for fracture predictions using approximate Bayesian inference methods



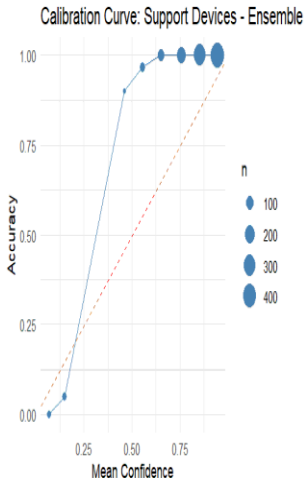
(a) MC Dropout



(b) Mean-Field VI

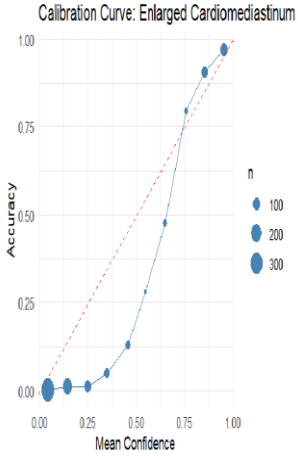


(c) SWAG

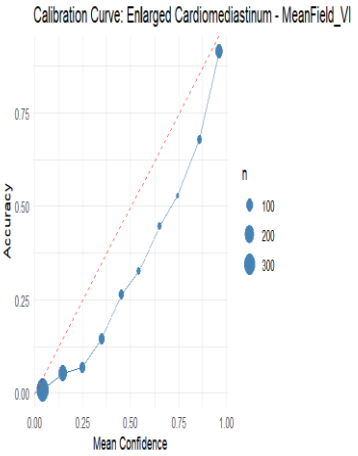


(d) Ensemble

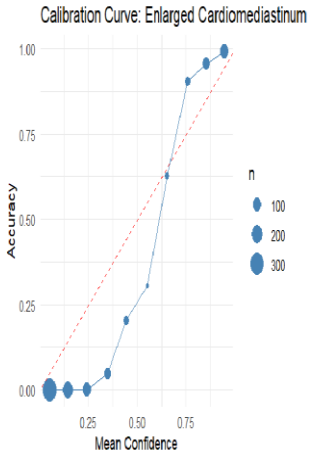
Figure 13: Reliability diagrams for support devices predictions using approximate Bayesian inference methods



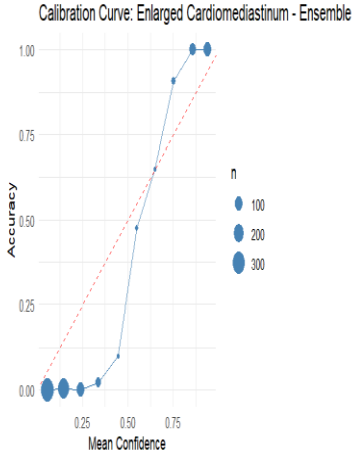
(a) MC Dropout



(b) Mean-Field VI



(c) SWAG



(d) Ensemble

Figure 14: Reliability (calibration) curves for Enlarged Cardiomediastinum across different Bayesian inference methods.

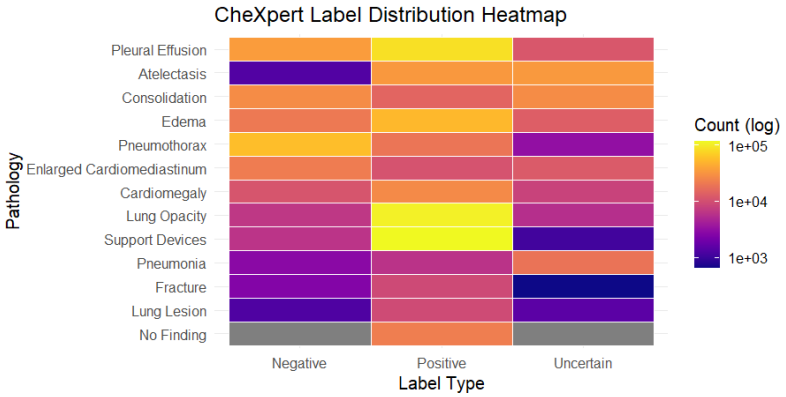


Figure 15: Heatmap of label distribution in the CheXpert training set across pathologies and label types.

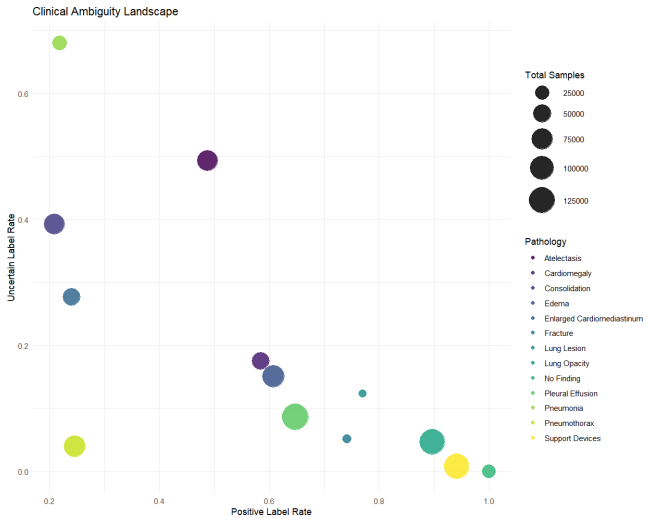
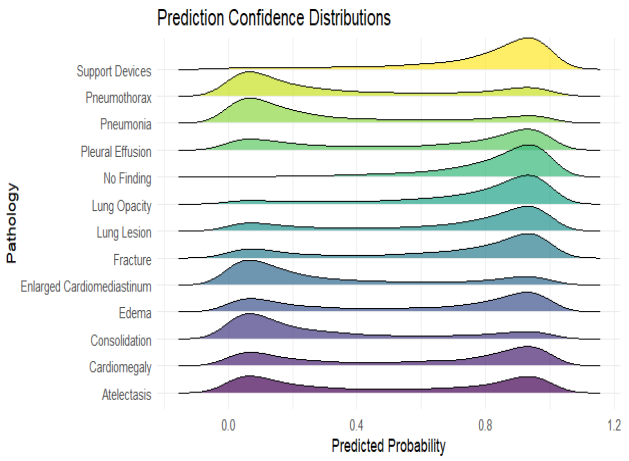
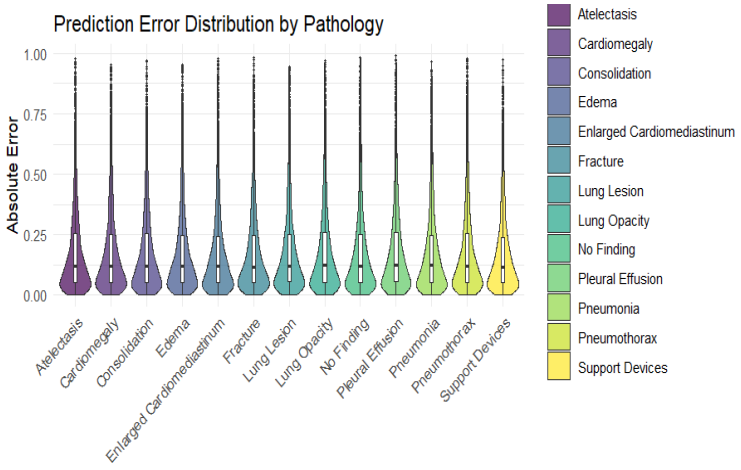


Figure 16: Clinical ambiguity landscape of CheXpert pathologies.



(a) Confidence Distributions



(b) Error Distributions

Figure 17: Analysis of model performance across pathologies: (a) ridge plots of predicted probability distributions and (b) violin plots of absolute prediction error.

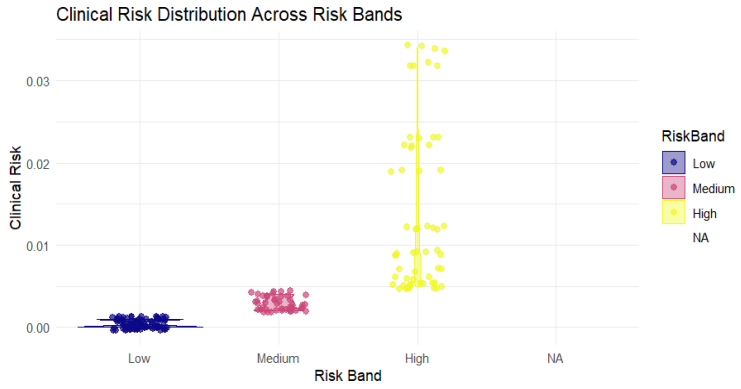


Figure 18: Distribution of clinical risk scores across categorized risk bands

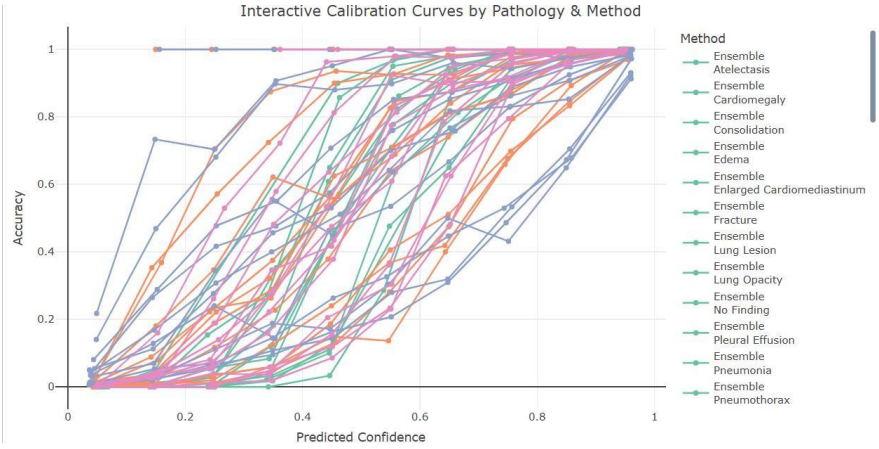


Figure 19: Interactive calibration curves for various pathologies and method

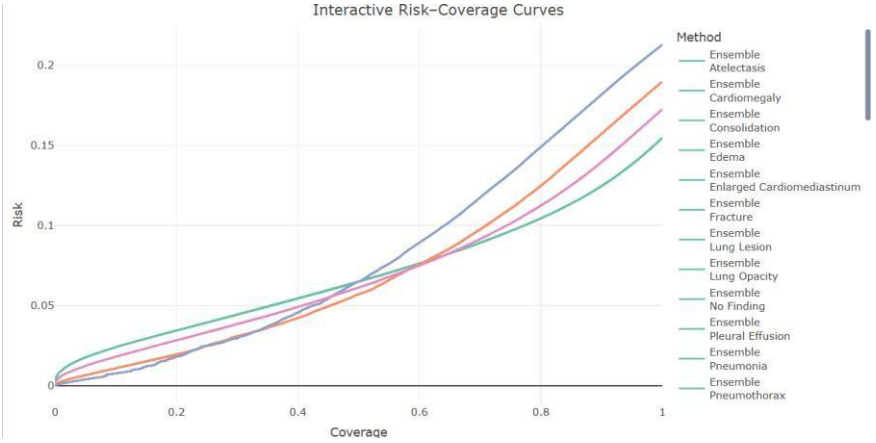


Figure 20: Interactive risk-coverage curves for multiple pathologies.

6 Conclusion

This study presents a comprehensive evaluation of Bayesian deep learning approaches for uncertainty-aware chest X-ray pathology detection, with a particular focus on calibration, reliability, and clinical risk sensitivity. Extensive experiments conducted on the CheXpert dataset demonstrate that deterministic deep neural networks and standard ensemble methods, while achieving high discriminative performance, often produce overconfident and poorly calibrated probability estimates—especially for clinically ambiguous and infrequent pathologies such as Pneumonia, Atelectasis, and Consolidation.

By integrating scalable approximate Bayesian inference techniques, including Monte Carlo Dropout, Mean-Field Variational Inference, and Stochastic Weight Averaging Gaussian, within a unified deep ensemble framework, the proposed approach effectively captures both intra-model epistemic uncertainty and inter-model diversity. Empirical results show that Bayesian deep ensembles consistently achieve the highest AUROC across thoracic pathologies, while Bayesian approximations—particularly Mean-Field Variational Inference—exhibit superior probabilistic calibration as reflected by lower Expected Calibration Error. These findings highlight an important trade-off between predictive discrimination and calibration, emphasizing the need to jointly assess accuracy and uncertainty in medical AI systems.

Furthermore, the analysis of label uncertainty, calibration behavior, and clinical risk ranking reveals that epistemic uncertainty strongly correlates with diagnostic ambiguity and class imbalance inherent in real-world chest radiograph datasets. Pathologies associated with high uncertainty rates and elevated calibration error naturally emerge as high-risk categories, underscoring the value of uncertainty-aware evaluation for selective prediction and risk-sensitive clinical triage.

In summary, this work demonstrates that explicit Bayesian uncertainty modeling is critical for the safe and reliable deployment of deep learning systems in chest X-ray interpretation. The proposed Bayesian deep ensemble framework provides a principled, scalable, and empirically effective solution for achieving well-calibrated predictions and clinically meaningful uncertainty estimates, thereby supporting robust human–AI collaboration in medical imaging.

7 Future Work

Several promising directions emerge from this study to further enhance uncertainty aware medical image analysis. First, future work will focus on extending the proposed Bayesian deep ensemble framework to explicitly model *aleatoric uncertainty* alongside epistemic

uncertainty, for example through heteroscedastic likelihoods or label-noise aware learning objectives, which may further improve robustness in the presence of ambiguous or low-quality annotations. Incorporating uncertainty-aware loss functions that jointly account for noisy and missing labels is particularly relevant for large-scale clinical datasets such as CheXpert.

Second, while this work primarily evaluates in-distribution performance, future studies should investigate robustness under distribution shift, including cross-institutional transfer, temporal dataset drift, and domain adaptation across different imaging devices and patient populations. Evaluating uncertainty estimates as indicators for out of distribution detection and selective referral to radiologists represents a critical step toward safe clinical deployment.

Third, extending the proposed framework to multi-task and multi-modal settings—such as joint learning across multiple radiological views or integration with clinical metadata, laboratory values, and electronic health records—may yield richer uncertainty representations and improved diagnostic performance. In addition, structured Bayesian priors informed by anatomical or pathophysiological knowledge could be explored to further enhance interpretability and clinical trust.

Finally, future work will emphasize prospective clinical validation and human–AI interaction studies to assess how calibrated uncertainty estimates influence radiologist decision-making, workflow efficiency, and patient outcomes. Integrating uncertainty aware predictions into interactive clinical decision support systems, including adaptive triage and risk-based prioritization, represents a crucial step toward the responsible and effective translation of Bayesian deep learning methods into real-world healthcare practice.

Data Availability

The CheXpert dataset used in this study is publicly available and can be accessed via the following link: https://www.kaggle.com/datasets/ashery/chexpert.

Code Availability

All the codes used for data preprocessing, statistical modeling, and figure generation is publicly available in a GitHub repository under an open-source license at: https://github.com/sthdas999/Bayesian-Deep-Learning-Approaches-for-Chest-X-Ray-Pathology-Detection .The repository includes annotated scripts and instructions to reproduce the full analysis pipeline.

Competing Interests

The authors declare that they have no competing interests.

Funding

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

Ethics Statement

This study did not involve human participants or animals requiring ethical approval. All data used were obtained from open-access healthcare experiments.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi and U. R. Acharya, “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges”, Information Fusion, vol. 76, (2021), pp. 243-297.

[2] E. Begoli, T. Bhattacharya and D. Kusnezov, “The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making”, Nature Machine Intelligence, vol. 1, no. 1, (2019), pp. 20-23.

[3] C. Blundell, J. Cornebise, K. Kavukcuoglu and D. Wierstra, “Weight Uncertainty in Neural Networks”, International Conference on Machine Learning (ICML), (2015), pp. 1613-1622.

[4] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”, International Conference on Machine Learning (ICML), (2016), pp. 1050-1059.

[5] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, “On Calibration of Modern Neural Networks”, International Conference on Machine Learning (ICML), (2017), pp. 1321-1330.

[6] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball and K. Shpanskaya, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison”, AAAI Conference on Artificial Intelligence, (2019), pp. 590-597.

[7] A. Kendall and Y. Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?”, Advances in Neural Information Processing Systems (NeurIPS), (2017).

[8] P. Lakhani and B. Sundaram, “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks”, Radiology, vol. 284, no. 2, (2017), pp. 574-582.

[9] B. Lakshminarayanan, A. Pritzel and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles”, Advances in Neural Information Processing Systems (NeurIPS), (2017), pp. 6402-6413.

[10] D. J. C. MacKay, “A Practical Bayesian Framework for Backpropagation Networks”, Neural Computation, vol. 4, no. 3, (1992), pp. 448-472.

[11] W. Maddox, T. Garipov, P. Izmailov, D. Vetrov and A. G. Wilson, “A Simple Baseline for Bayesian Uncertainty in Deep Learning”, Advances in Neural Information Processing Systems (NeurIPS), (2019), pp. 13153-13164.

[12] R. M. Neal, “Bayesian Learning for Neural Networks”, Springer, vol. 118, (2012).

[13] L. Oakden-Rayner, J. Dunnmon, G. Carneiro and C. Ré, “Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging”, Proceedings of the ACM Conference on Health, Inference, and Learning, (2020), pp. 151-159.

[14] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul and C. P. Langlotz, “Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists”, PLoS Medicine, vol. 15, no. 11, (2018), e1002686.

[15] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz and K. Shpanskaya, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”, arXiv Preprint arXiv:1711.05225, (2017).

[16] H. Ritter, A. Botev and D. Barber, “A Scalable Laplace Approximation for Neural Networks”, International Conference on Learning Representations (ICLR), (2018).

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, “ChestX-ray14: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017), pp. 2097-2106.

[18] A. G. Wilson and P. Izmailov, “Bayesian Deep Learning and a Probabilistic Perspective of Generalization”, Advances in Neural Information Processing Systems (NeurIPS), (2020).