Data-Driven Priors for Uncertainty-Aware Deterioration Risk Prediction with Multimodal Data

Anonymous authors
Paper under double-blind review

Abstract

Safe predictions are a crucial requirement for integrating predictive models into clinical decision support systems. One approach for ensuring trustworthiness is to enable the models' ability to express their uncertainty about individual predictions. However, current machine learning models frequently lack reliable uncertainty estimation, hindering real-world deployment. This is further observed in multimodal settings, where the goal is to enable effective information fusion. In this work, we propose `MedCertAIn`, a predictive uncertainty framework that leverages multimodal clinical data for in-hospital risk prediction to improve model performance and reliability. We design data-driven priors over neural network parameters using a hybrid strategy that considers cross-modal similarity in self-supervised latent representations and modality-specific data corruptions. We train and evaluate the models with such priors using clinical time-series and chest X-ray images from the publicly-available datasets MIMIC-IV and MIMIC-CXR. Our results show that `MedCertAIn` significantly improves predictive performance and uncertainty quantification compared to state-of-the-art deterministic baselines and alternative Bayesian methods. These findings highlight the promise of data-driven priors in advancing robust, uncertainty-aware AI tools for high-stakes clinical applications.

1 Introduction

Trustworthy implementation of machine learning models in healthcare requires robust uncertainty measurements (Begoli et al., 2019; Gruber et al., 2023), considering the safety-critical nature of clinical practice. Uncertainty Quantification (UQ) approaches provide an essential layer of reliability in addition to accurate machine learning models for improved and safer decision-making (Nemani et al., 2023). Uncertainty in machine learning models can be due to model parameters, noise and bias of the calibration data, or deployment of the model in an out-of-distribution scenario (Miller et al., 2014). All of these are inherent characteristics of any real-world clinical scenario. Artificial Intelligence (AI) systems that reliably provide uncertainty metrics can enhance the confidence of healthcare professionals when using these systems and, at the same time, improve the effectiveness of the predictions (Seoni et al., 2023) .

Nonetheless, the study of UQ within the area of machine learning for healthcare has been limited (Kompa et al., 2021; Lechuga et al., 2025). A comprehensive understanding of the uncertainty present in healthcare applications and how to address it is still fragmented (Han et al., 2011), which may be due to the limited underlying theory on how to best adapt predictive uncertainty for clinical tasks (Begoli et al., 2019). Other reasons include the complexity of scaling UQ in real-time clinical systems, limited empirical evaluation of different methods due to the lack of well-constructed priors by medical experts (Zou et al., 2023), and the high prevalence of data shifts in real-world clinical applications. All of these can negatively affect predictive performance (Ovadia et al., 2019b; Xia et al., 2022), further emphasizing the need for better measurements of uncertainty in predictive models.

In addition, developing machine learning systems that are suitable for healthcare applications require the use of different data modalities simultaneously, to reflect the multimodal nature of the human decision-making process (Saab et al., 2024). However, existing work on UQ in healthcare has been mainly studied in the

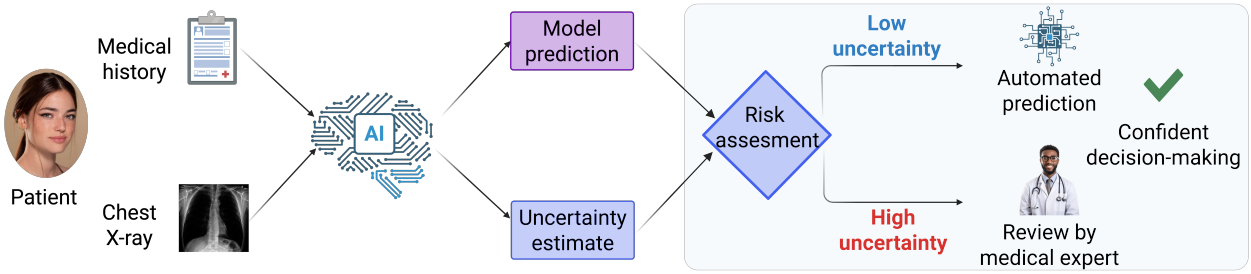

Figure 1: **Uncertainty-Aware Clinical Decision-making.** Patient multimodal data is processed by our uncertainty-aware model to produce a prediction and an uncertainty estimate. Clinicians accept confident predictions or defer uncertain cases for further review, improving workflow efficiency and supporting reliable decision-making.

unimodal setting, with a particular focus on medical imaging applications (Gawlikowski et al., 2021). Hence, effective quantification of predictive uncertainty in the context of multimodal clinical problems remains a challenging and unsolved task (Tran et al., 2022).

To address these gaps, we introduce `MedCertAIn`, a multimodal uncertainty-aware framework for deterioration risk prediction that explicitly recognizes and defers low-confidence cases for human review, which is considered a key requirement for real-world clinical decision support (Figure 1). Our method combines Bayesian learning with automated, label-free priors to enable selective prediction without manual annotation or domain-specific expert input, making it scalable across diverse clinical settings. We summarize our key contributions as follows:

1. We propose `MedCertAIn`, a multimodal uncertainty-aware framework that combines Bayesian learning and variational inference to augment existing architectures with principled uncertainty estimates and improved reliability.
2. We design a flexible label-free pipeline for automatically constructing multimodal priors by considering cross-modal similarity in self-supervised latent representations and modality-specific input perturbations, eliminating the need for expert-annotated high-uncertainty context sets.
3. We train and evaluate our framework on publicly available multimodal data comprising clinical time-series and chest X-ray images from MIMIC-IV (Johnson et al., 2021) and MIMIC-CXR (Johnson et al., 2019). Our findings demonstrate improved predictive performance and reliability.
4. We implement our framework in `JAX` and make the code publicly available to promote future research in uncertainty-aware multimodal clinical modeling: `https://anonymous.4open.science/r/medcertain_tmlr-8154`.

The paper is organized as follows: Section 2 provides an overview of related work on multimodal learning and UQ in healthcare, Section 3 presents the proposed `MedCertAIn` framework and training methodology, Section 4 describes the experimental setting, Section 5 presents the results and empirical findings, and Section 6 discusses the implications, limitations, and future directions.

2 Related Work

2.1 Multimodal Learning in Healthcare

Multimodal learning approaches seek to leverage complementary information from different data modalities to enhance the models' predictive capabilities (Hayat et al., 2022). In healthcare, this paradigm is particularly relevant due to the heterogeneous nature of clinical data, which often includes structured electronic health records, medical imaging, physiological signals, and free-text reports. Many approaches for combining information across modalities exist, with multimodal fusion being the most widely adopted strategy (Huang et al., 2020; Hayat et al., 2022).

Early work primarily focused on integrating multiple imaging modalities or views for tasks such as segmentation and disease characterization, particularly in neuroimaging (Zhang et al., 2020; Calhoun & Sui, 2016). More recent studies have extended multimodal learning to combine fundamentally different data types, such as imaging with time-series or tabular clinical data, demonstrating improved performance in prognostic and risk prediction tasks (Muhammad et al., 2021). This trend is exemplified by multimodal systems developed for patient deterioration and outcome prediction, including COVID-19 prognosis using combined imaging and clinical variables (Shamout et al., 2021; Jiao et al., 2021). While multimodal integration enables richer patient representations, it also introduces additional sources of uncertainty arising from heterogeneity across modalities, imperfect alignment in time and semantics, and potential disagreement between modality-specific signals (Azizmalayeri et al., 2025). These challenges are inherent of real-world scenarios motivating the need for modeling approaches that explicitly account for uncertainty in multimodal settings (Kurz et al., 2022).

2.2 Uncertainty Quantification in Medical Applications

The safety-critical nature of clinical decision-making has motivated growing interest in UQ for machine learning models in healthcare (Gawlikowski et al., 2021; Seoni et al., 2023). Most existing work on UQ in medical applications has focused on unimodal settings, particularly medical imaging, where uncertainty estimates have been used to support tasks such as brain tumor segmentation (Jungo et al., 2018), skin lesion classification (DeVries & Taylor, 2018), and diabetic retinopathy detection (Filos et al., 2019; Band et al., 2021; Nado et al., 2022). These studies demonstrate the potential of UQ to improve interpretability and model trust by distinguishing confident predictions from ambiguous cases.

However, uncertainty in real-world healthcare systems arises from multiple sources, including data noise, dataset shift, model misspecification, and deployment in out-of-distribution clinical environments (Lechuga et al., 2025). Existing UQ approaches rarely account for these factors jointly, nor do they explicitly address the challenges introduced by multimodal clinical data, particularly those designed to support selective prediction and deferral in deployment settings, where uncertainty may stem from modality disagreement or missing and corrupted inputs (Liang et al., 2023).

2.3 Stochastic Neural Networks and Variational Inference

Stochastic neural networks extend deterministic neural networks by treating their parameters as random variables rather than fixed point estimates. This probabilistic formulation enables the model to capture epistemic uncertainty and provides a principled foundation for Bayesian inference over network weights. Hence, for a given neural network $f(\cdot)$, a stochastic version $f(\cdot\,;\Theta)$ is defined in terms of stochastic parameters Θ. For an observation model $p_{Y|X,\Theta}$ and a prior distribution over parameters p_Θ, Bayesian inference provides a principled framework for modeling predictive uncertainty by inferring the posterior distribution over parameters given the observed data, $p_{\Theta|\mathcal{D}}$ (MacKay, 1992; Neal, 1996). However, since neural networks are non-linear in their parameters, exact inference over the stochastic network parameters is analytically intractable. Variational inference is an approach that seeks to avoid this intractability by framing posterior inference as finding an approximation q_Θ to the posterior $p_{\Theta|\mathcal{D}}$ via the variational optimization problem:

$$\min_{q_\Theta \in \mathcal{Q}_\Theta} \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta|\mathcal{D}}) \Longleftrightarrow \max_{q_\Theta \in \mathcal{Q}_\Theta} \mathcal{F}(q_\Theta)$$

where $\mathcal{F}(q_\Theta)$ is the variational objective:

$$\mathcal{F}(q_\Theta) \doteq \mathbb{E}_{q_\Theta}[\log p(y_\mathcal{D} \,|\, x_\mathcal{D}, \Theta)] - \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_\Theta),$$

$\mathcal{Q}_\Theta$ is a variational family of distributions (Wainwright & Jordan, 2008), $\mathbb{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence (Kullback & Leibler, 1951), and $(x_\mathcal{D}, y_\mathcal{D})$ are the training data. One particularly simple type of variational inference is Gaussian mean-field variational inference (Blundell et al., 2015; Graves, 2011), where the posterior distribution over network parameters is approximated by a Gaussian distribution with a diagonal covariance matrix. This method enables stochastic optimization and can be scaled to large neural networks (Hoffman et al., 2013). However, Gaussian mean-field variational inference has been shown to underperform with deterministic neural networks when uninformative, standard Gaussian priors are used (Ovadia et al., 2019a; Rudner et al., 2022a).

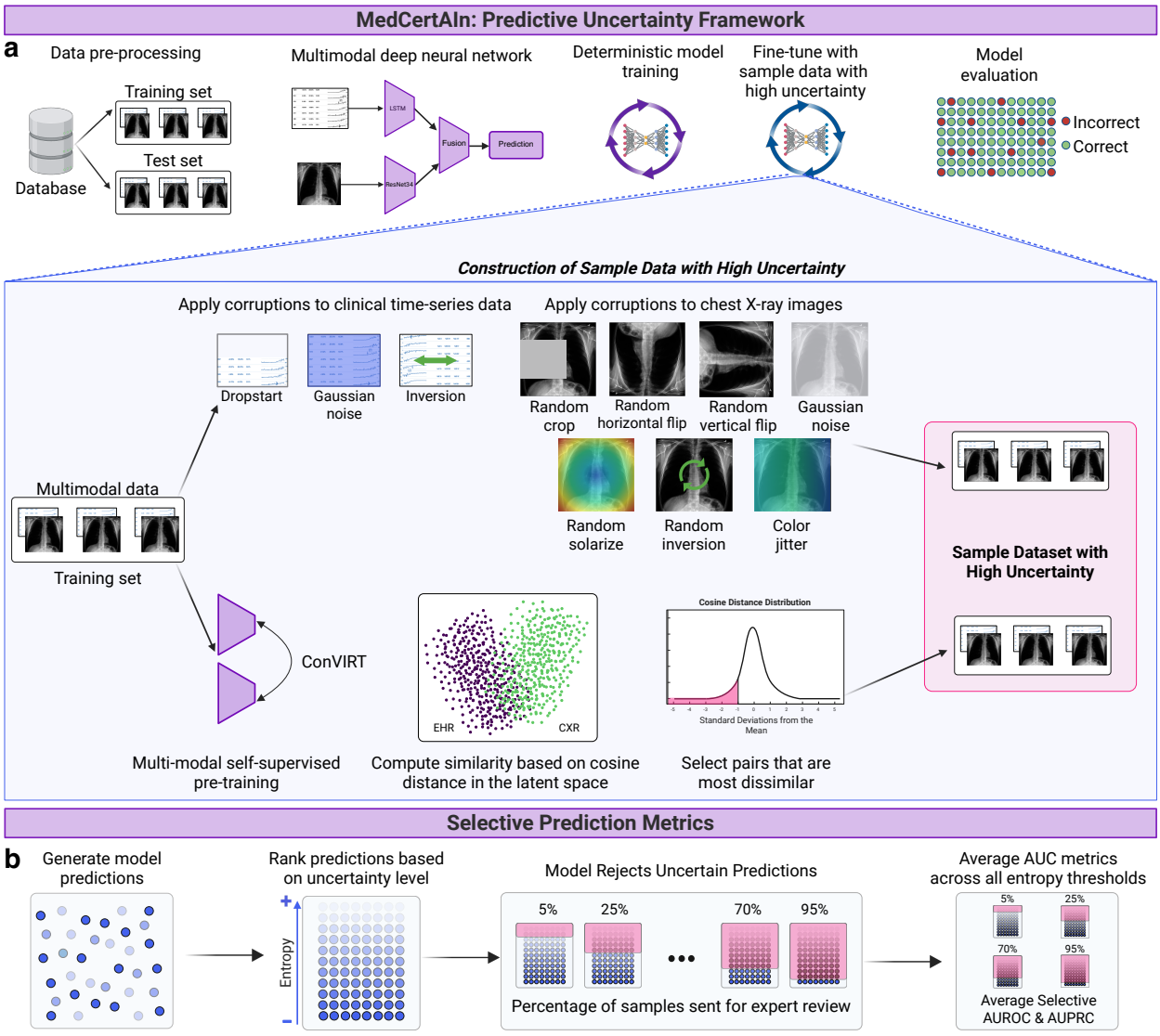

Figure 2: **`MedCertAIn`: A Predictive Uncertainty Framework a)** We construct paired multimodal samples for training/evaluation and generate high-uncertainty context data. From the training set, we create corrupted pairs via noise and transformations, and additionally pre-train a self-supervised model to select low-similarity pairs in latent space, reflecting high cross-modal uncertainty. We combine these context sets and fine-tune the model using Bayesian learning, evaluating with standard ROC and selective prediction metrics. **b) Selective Prediction Metrics:** We rank predictions by uncertainty (Shannon entropy) and evaluate performance on retained subsets across coverage thresholds (0–100%). Averaging metrics across thresholds yields selective AUROC/AUPRC, enabling clinicians to choose a deferral point where the model is least reliable.

3 Methods

`MedCertAIn` is a multimodal uncertainty-aware learning framework designed to produce calibrated predictions and to explicitly defer low-confidence cases in safety-critical clinical settings. Our contribution is an uncertainty-aware training objective that augments standard multimodal prediction with principled uncertainty estimation, while remaining independent of the underlying fusion architecture. Rather than modifying

network design, `MedCertAIn` operates at the level of context priors selection and loss formulation, enabling reliable selective prediction across existing multimodal models.

3.1 Learning Objective

We consider a supervised multimodal fusion task on a dataset $\mathcal{D} \doteq \{(x_n^{ehr}, x_n^{cxr}, y_n)\}_{n=1}^{N} = (X_{\mathcal{D}}^{ehr}, X_{\mathcal{D}}^{cxr}, Y_{\mathcal{D}})$, where each sample consists of paired electronic health record time-series and chest X-ray images. As illustrated in Figure 2.a, the two modalities are processed by modality-specific encoders Φ^{ehr} and Φ^{cxr}, respectively. The resulting feature representations are concatenated and passed to a classifier $g(\cdot)$ to produce a fusion prediction $\hat{y}$, which is optimized using supervised learning with respect to the ground truth labels $y \in \{0, 1\}$.

To mitigate the limitations of variational inference under non-informative priors, we use information from our two input modalities to construct data-driven priors that can help find an approximate posterior distribution with desirable properties (e.g., an induced predictive distribution with reliable uncertainty-aware estimation), obtained using variational inference approximation (Graves, 2011). Specifically, we construct this data-driven prior over some set of model parameters Ψ and condition it on a set of context points X_c with corresponding labels Y_c, that is $p(\psi|x_c)$. To construct a meaningful prior, we need to specify a distribution over the set of context points, p_{x_c}, which we refer to as sample data with high uncertainty. Used in combination with the training set, the context set is used as input during training to help the model learn to separate data points it struggles most to classify. Hence, this sample data with high uncertainty encompasses distributionally shifted points, where we want the model's uncertainty to be higher, which can then be used to guide further clinical decision-making.

For training our stochastic model, we extend the variational objective with an uncertainty regularization term that makes use of the set of context points:

$$\mathcal{F}(q_\Theta) \doteq \underbrace{\mathbb{E}_{q_\Theta}[\log p(y_{\mathcal{D}} \,|\, x_{\mathcal{D}}, \Theta; f)]}_{\text{Expected log-likelihood}} - \underbrace{\mathbb{D}_{\mathrm{KL}}(q_\Theta \,\|\, p_\Theta)}_{\text{KL regularization}} + \underbrace{\mathbb{E}_{q_\Theta}[\mathbb{E}_{p_{X_c,Y_c}}[\log \tilde{p}(Y_c \,|\, X_c, \Theta;)]]}_{\text{Uncertainty regularization}}$$

Letting $p_{Y_c|X_c}(y_c \,|\, x_c) = \delta(\mathbf{0})$ to encourage high uncertainty in the predictions on the set of context points, where $\delta(\cdot)$ is the Dirac delta function, we obtain the simplified objective:

$$\mathcal{F}(q_\Theta) \doteq \underbrace{\mathbb{E}_{q_\Theta}[\log p(y_{\mathcal{D}} \,|\, x_{\mathcal{D}}, \Theta; f)]}_{\text{Expected log-likelihood}} - \underbrace{\mathbb{D}_{\mathrm{KL}}(q_\Theta \,\|\, p_\Theta)}_{\text{KL regularization}} + \underbrace{\mathbb{E}_{q_\Theta}[\mathbb{E}_{p_{X_c}}[\log \tilde{p}(\mathbf{0} \,|\, X_c, \Theta; f)]]}_{\text{Uncertainty regularization}}$$

In practice, each training step performs a forward pass on the combined training and context sets. We optimize the expected log-likelihood via binary cross-entropy between the prior distribution and the approximate posterior distribution, and compute the KL and uncertainty regularization terms using Monte Carlo estimation. Gradients are obtained via reparameterization of gradients for mean-field variational inference (Blundell et al., 2015). Additional derivations of the tractable objective under the proposed context-conditioned prior are provided in Appendix C.

3.2 Constructing the Context Set in A Priori High-Uncertainty Regions

We define a context set (X_c, Y_c) drawn from a distribution p_{X_c,Y_c}, representing samples in a priori high-uncertainty regions (i.e., distributionally shifted or ambiguous points). From the multimodal training set $(X_{\mathcal{D}}^{ehr}, X_{\mathcal{D}}^{cxr})$, we construct a context set $(X_{\mathcal{C}}^{ehr}, X_{\mathcal{C}}^{cxr})$ using two label-free strategies which enable data-driven prior construction with minimal human intervention: (1) applying controlled corruptions to both modalities, and (2) selecting samples with low cross-modal similarity in a self-supervised latent space. The latter is motivated by the assumption that modality mismatch corresponds to higher predictive uncertainty.

3.2.1 Modality-specific Data Perturbations

We induce a controlled distribution shift by applying data corruptions to the original training set. For EHR time-series, we construct $X_{\mathcal{C}}^{ehr}$ by: (i) dropping an initial segment of the sequence up to a threshold, (ii)

adding Gaussian noise, and (iii) reversing the time dimension. For chest X-rays, we construct $X_{\mathcal{C}}^{cxr}$ using seven perturbations: random crop, horizontal/vertical flip, Gaussian blur, solarize, invert, and color jitter. Corruption magnitudes are fixed during training using commonly adopted augmentation settings.

3.2.2 Cross-Modal Similarity in Multimodal Latent Space

We also construct a context set by selecting samples whose modality representations are dissimilar in a self-supervised latent space. We obtain per-modality representations $(\Phi^{\text{ehr}}, \Phi^{\text{cxr}})$ using a ConVIRT model trained with the infoCNE loss (Zhang et al., 2022; Hayat et al., 2022). We compute cosine distance and select samples falling below a threshold defined as a left-tail cutoff of the distance distribution.

Inter-modal similarity. We compute the mean cross-modal cosine distance across n training samples and define:

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n} \cos(\Phi_i^{\text{ehr}}, \Phi_i^{\text{cxr}}), \qquad t = \gamma_1 - v \cdot \sigma,$$

where σ is the standard deviation of distances over the training set and $v \in [1, 2]$. We include samples satisfying $\gamma_1 < t$:

$$X_{\mathcal{C}}^{ehr} = \{X_c^{\text{ehr}} : x^{\text{ehr}} \cdot \mathbb{1}_{\gamma_1 < t}\}, \qquad X_{\mathcal{C}}^{cxr} = \{X_c^{\text{cxr}} : x^{\text{cxr}} \cdot \mathbb{1}_{\gamma_1 < t}\}.$$

Inter- and intra-modal similarity. We additionally incorporate intra-modal distances to the mean latent representation of each modality:

$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{n} \cos(\Phi_i^{\text{ehr}}, \overline{\Phi}_{\text{ehr}}), \qquad \gamma_3 = \frac{1}{n}\sum_{i=1}^{n} \cos(\Phi_i^{\text{cxr}}, \overline{\Phi}_{\text{cxr}}),$$

and aggregate them with the inter-modal distance:

$$\gamma_4 = \frac{\gamma_1 + \gamma_2 + \gamma_3}{3}, \qquad t_4 = \gamma_4 - c \cdot \sigma_4,$$

where σ_4 is the standard deviation of γ_4 over the training set. We then select samples satisfying $\gamma_4 < t_4$:

$$X_{\mathcal{C}}^{ehr} = \{X_c^{\text{ehr}} : x^{\text{ehr}} \cdot \mathbb{1}_{\gamma_4 < t_4}\}, \qquad X_{\mathcal{C}}^{ehr} = \{X_c^{\text{cxr}} : x^{\text{cxr}} \cdot \mathbb{1}_{\gamma_4 < t_4}\}.$$

Finally, for training `MedCertAIn`, the high-uncertainty sample set is constructed by combining the corrupted training examples and the dissimilar samples identified via self-supervised latent-space selection.

3.3 Uncertainty Measurement

To quantify predictive uncertainty, we compute uncertainty scores individually for each test sample based on the predictive distribution of the model. For a given input x, the stochastic network induces a predictive distribution $p(y \mid x)$, obtained via Monte Carlo sampling from the variational posterior. We quantify uncertainty using the Shannon entropy of the predictive distribution:

$$\mathcal{H}(p(y \mid x)) = -\sum_{c \in \{0,1\}} p(y = c \mid x) \log p(y = c \mid x), \tag{1}$$

where higher entropy indicates greater predictive uncertainty.

Shannon entropy provides a simple and widely used proxy for predictive uncertainty in Bayesian neural networks (Rudner et al., 2023a;b). In our setting, entropy scores are computed per data point and used to rank predictions for selective prediction which evaluates the models skill at detecting points that the model would recommend for further review, enabling uncertainty-aware clinical decision support.

4 Experimental Setting

4.1 Dataset

To train and evaluate our framework, we used two publicly available modalities: EHR time-series data from MIMIC-IV (Johnson et al., 2021) and chest X-ray images from MIMIC-CXR (Johnson et al., 2019). MIMIC-IV contains more than 60,000 ICU stays collected between 2001 and 2019. We only used complete paired samples, such that each example includes both modalities with no missing data. We focus on in-hospital mortality prediction, a clinically relevant task for ICU decision support systems (Sadeghi et al., 2018; Awad et al., 2017), to support clinical decision-making and resource allocation by identifying patients at higher risk of deterioration, enabling timely intervention (Rajpurkar et al., 2022). Mortality prediction is a binary classification task that estimates in-hospital mortality after 48 hours in the ICU (Hayat et al., 2022). We excluded ICU stays shorter than 48 hours, resulting in 6,215 paired patient samples. We used an 80/20 train-test split and performed five-fold cross-validation on the training set with a 70/10 train-validation ratio. All splits are disjoint at the patient level. We follow Hayat et al. (2022) to pre-process and align the clinical time-series data and chest X-ray images for this task. Table 1 summarizes the multimodal sample sizes across training, validation, test and context set splits used for `MedCertAIn`.

Table 1: **Dataset Summary**: Summary of multimodal data samples used during training for the in-hospital mortality task. Only our proposed method `MedCertAIn` makes use of the high-uncertainty set (5,347 samples) which is constructed by combining corrupted training examples with 862 additional samples selected via cross-modal similarity in the self-supervised latent space.

Data split \ Model	Baseline Models	`MedCertAIn`
Training	4,485	4,485
Validation	488	488
Test	1,242	1,242
High Uncertainty Set	–	5,347

4.2 Multimodal Backbone Neural Network

We adopt MedFuse (Hayat et al., 2022) as the backbone neural network architecture, a simple and robust fusion module for multimodal clinical prediction using paired EHR time-series and imaging data. MedFuse follows an early-fusion design, where modality-specific feature extractors encode each input before combining their representations for downstream prediction. Given a paired multimodal sample $(x^{ehr}, x^{cxr}) \in \mathcal{D}$, Φ_{ehr} encodes EHR time-series using a two-layer long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997), which captures temporal dependencies in sequential clinical measurements. Chest X-ray images are encoded using a ResNet-34 convolutional neural network Φ_{cxr} (He et al., 2015), which extracts high-level spatial and semantic features from medical images. The resulting modality-specific representations are concatenated and passed to a classifier $g(\cdot)$ with a sigmoid activation to produce the predicted mortality risk $\hat{y}$.

4.3 Model Baselines

We compare `MedCertAIn` against three primary baselines. First, we use MedFuse (Hayat et al., 2022), our deterministic multimodal backbone, which serves as the architectural baseline without uncertainty-aware training. Second, we evaluate DrFuse (Yao et al., 2024), a transformer-based multimodal model that disentangles shared and modality-specific representations and dynamically weights EHR and imaging inputs. We adapt DrFuse from clinical condition classification to the in-hospital mortality prediction task. Finally, we include a stochastic baseline based on Gaussian mean-field variational inference with an uninformative prior (Blundell et al., 2015; Graves, 2011), isolating the effect of Bayesian learning without informative, data-driven priors. All baselines are evaluated under identical data splits and experimental protocols to ensure fair comparison.

4.4 Evaluation Metrics

We evaluate test performance using AUROC and AUPRC. To assess predictive uncertainty, we additionally compute selective prediction metrics. Selective prediction introduces a "reject option" $\perp$ via a selection function $s : \mathcal{X} \to \mathbb{R}$ that determines whether to output a prediction for an input $x \in \mathcal{X}$ (El-Yaniv et al., 2010). Given a rejection threshold τ and using entropy as the selection score, the prediction rule is:

$$(p(y \mid \cdot, \theta; f), s)(x) = \begin{cases} p(y \mid x, \theta; f), & \text{if } s \leq \tau \\ \perp, & \text{otherwise.} \end{cases}$$

We compute AUROC and AUPRC over rejection thresholds $\tau = 0\%, \ldots, 99\%$ and report the average across thresholds as selective AUROC and selective AUPRC. Selective prediction metrics quantify predictive performance together with a model's uncertainty, providing an intuitive measure of reliability. In practice, they offer a simple mechanism to flag uncertain predictions for clinical review, improving interpretability for clinicians while ensuring that only low-confidence cases are deferred, optimizing clinical resources.

4.5 Hyperparameter Tuning and Model Selection

4.5.1 Deterministic models

For each task, we run 50 hyperparameter configurations, each evaluated over five runs with different random seeds (250 runs total). For every run, we randomly sample train/validation splits and initialize model weights. We sample the learning rate uniformly from $[10^{-5}, 10^{-2}]$, select the regularizer scale from $\{0, 0.1, 1, 10, 100\}$, and the number of training epochs from $\{5, 10, 15, 20, 30\}$. We use a fixed batch size of 16 and cosine decay with $\alpha = 0$. For each run, we save the checkpoint from the final training epoch and select the configuration with the highest average validation AUROC across the five seeds. Using the selected configuration, we retrain on the combined training and validation data and report test performance averaged over five random initializations. We compute standard errors across these five final runs.

4.5.2 Stochastic models

We initialize stochastic models from the five best deterministic checkpoints and perform hyperparameter tuning using the same protocol (50 configurations $\times$ 5 seeds; 250 runs per task). The learning rate is sampled one order of magnitude above and below the best deterministic learning rate, and training epochs are sampled from $\{5, 10, 15, 20, 25, 30\}$. We use batch size 16, cosine decay with $\alpha = 0$, and sample the context batch size from $\{16, 32\}$ due to the additional context data used for Bayesian learning. As stochastic models include mean and variance parameters and require forward passes over sampled context points, fine-tuning is more computationally expensive than the deterministic setting. Table D.3 summarizes the stochastic search space. With the optimal configuration, we train on the full training plus validation data and evaluate five stochastic models with different random initializations (5 runs). This procedure is repeated for each stochastic variant trained with different high-uncertainty context sets.

All models are trained with Adam and evaluated on the test set using means and standard errors over five seeds. Experiments were run on NVIDIA A100 and V100 80GB Tensor Core GPUs.

5 Results

5.1 `MedCertAIn` Achieves State-Of-The-Art Performance and Reliable Selective Prediction

Table 2 reports in-hospital mortality prediction results for the deterministic baselines DrFuse (Yao et al., 2024) and MedFuse (Hayat et al., 2022), a Bayesian neural network with a naive Gaussian prior (Blundell et al., 2015; Rudner et al., 2022a), and our proposed method `MedCertAIn`. Overall, `MedCertAIn` achieves the strongest standard predictive performance, with AUROC 0.835 ± 0.001 and AUPRC 0.498 ± 0.002. Relative to MedFuse, this corresponds to a 1.93% gain in AUROC and a 7.00% gain in AUPRC, indicating improved

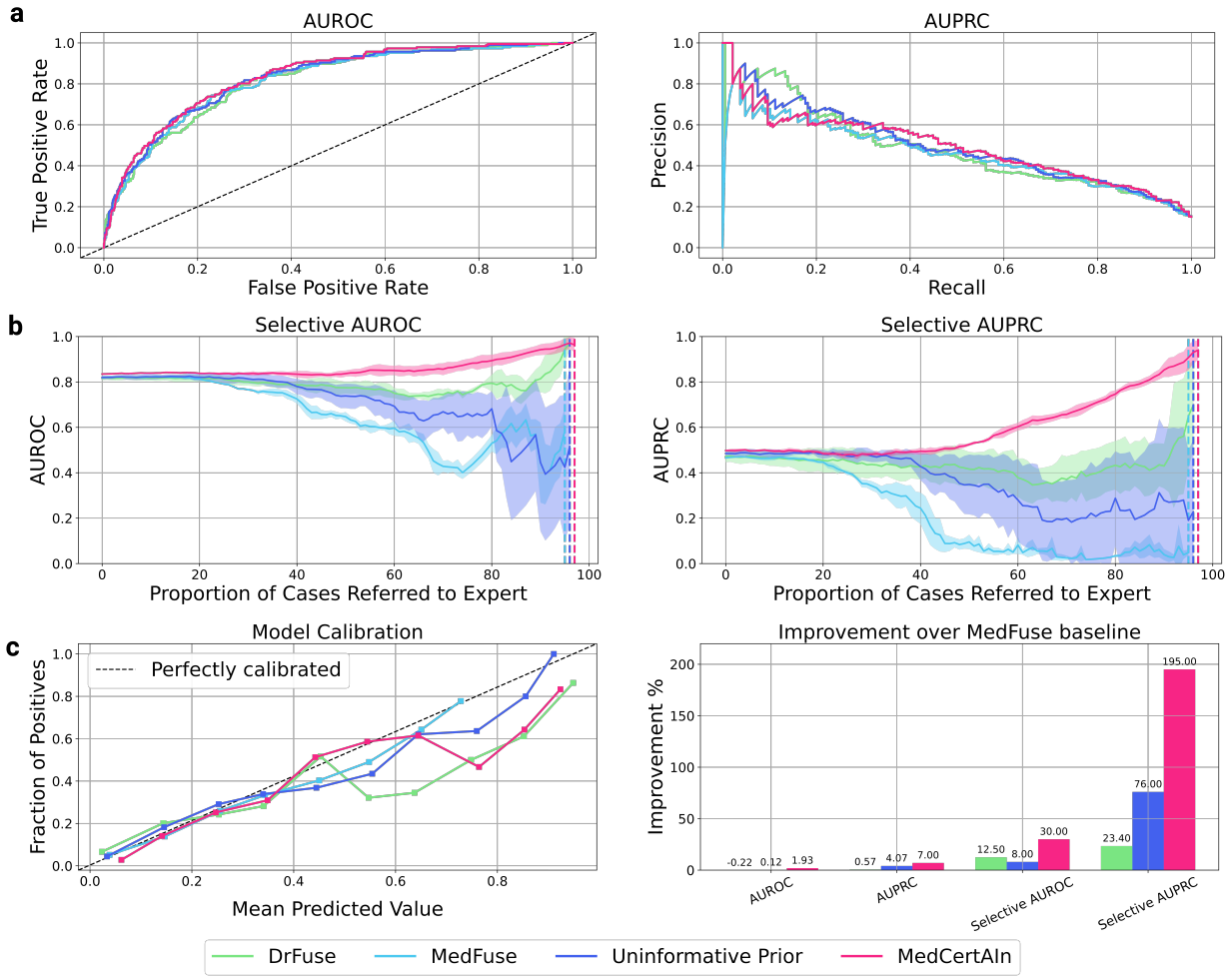

Figure 3: **`MedCertAIn` Achieves State-Of-The-Art Performance and Reliable Selective Prediction. a)** In standard metrics, the differences across baselines are very small for predictive performance. **b)** Analyzing the selective prediction metrics trends, we observe larger variance for each baseline. In particular, `MedCertAIn` shows the most stable signal across seeds, showing an enhanced capability of detecting high-uncertainty patients. **c)** The difference between `MedCertAIn` and the baseline MedFuse is even more significant in selective prediction metrics, with calibration scores similar across baselines.

discrimination and stronger performance under class imbalance, which is critical for reliable mortality prediction.

Beyond standard metrics, selective prediction is essential in clinical workflows, enabling models to flag uncertain cases for clinician review and improving safety under deployment. `MedCertAIn` consistently improves selective performance, achieving selective AUROC 0.857 ± 0.005 and selective AUPRC 0.599 ± 0.002. Compared to MedFuse, this yields a 30.0% improvement in selective AUROC and a 195% increase in selective AUPRC. Plots in Figure 3 show that our method is the most stable and performing across baselines for selective prediction. `MedCertAIn` provides both higher predictive performance and more reliable uncertainty estimates, facilitating clinician–AI cooperation by deferring cases that are more likely to be misclassified and supporting more efficient allocation of clinical resources.

5.2 Comparison Across Informative Multimodal Data-Driven Priors

We perform ablations over high-uncertainty sample selection strategies to quantify the impact of our data-driven priors. Table 3 compares a naive Gaussian prior, random corruptions, inter-modal similarity, inter-

Table 2: `MedCertAIn` **Achieves State-Of-The-Art Performance and Reliable Selective Prediction**: Quantitative performance of the deterministic and naive stochastic baselines compared to our proposed method `MedCertAIn` which consistently beats all other models with significant improvement in selective prediction metrics.

Model \ Metric	AUROC	AUPRC	Selective AUROC	Selective AUPRC
DrFuse (Yao et al., 2024)	$0.817_{\pm 0.005}$	$0.472_{\pm 0.013}$	$0.785_{\pm 0.007}$	$0.437_{\pm 0.035}$
MedFuse (Hayat et al., 2022)	$0.819_{\pm 0.000}$	$0.466_{\pm 0.002}$	$0.659_{\pm 0.005}$	$0.203_{\pm 0.006}$
Uninformative Prior (Rudner et al., 2022a)	$0.820_{\pm 0.004}$	$0.485_{\pm 0.006}$	$0.711_{\pm 0.034}$	$0.357_{\pm 0.058}$
`MedCertAIn` (**Ours**)	$\mathbf{0.835}_{\pm 0.001}$	$\mathbf{0.498}_{\pm 0.002}$	$\mathbf{0.857}_{\pm 0.005}$	$\mathbf{0.599}_{\pm 0.002}$
Improvement (%)	1.93%	7.00 %	30.0%	195%

Table 3: **Comparison Across Informative Multimodal Data-Driven Priors.** Quantitative performance across metrics for all different model ablations of data samples with high uncertainty.

Ablation	AUROC	AUPRC	Selective AUROC	Selective AUPRC
Uninformative Prior	$0.820_{\pm 0.004}$	$0.485_{\pm 0.006}$	$0.711_{\pm 0.034}$	$0.357_{\pm 0.058}$
Random Corruptions	$0.832_{\pm 0.001}$	$0.494_{\pm 0.002}$	$0.841_{\pm 0.006}$	$0.579_{\pm 0.011}$
Inter Similarity	$0.821_{\pm 0.001}$	$0.470_{\pm 0.002}$	$0.749_{\pm 0.013}$	$0.423_{\pm 0.032}$
Inter-Intra Similarity	$0.821_{\pm 0.001}$	$0.470_{\pm 0.002}$	$0.755_{\pm 0.007}$	$0.434_{\pm 0.020}$
`MedCertAIn I`	$\mathbf{0.835}_{\pm 0.001}$	$\mathbf{0.498}_{\pm 0.002}$	$\mathbf{0.857}_{\pm 0.005}$	$\mathbf{0.599}_{\pm 0.002}$
`MedCertAIn II`	$0.834_{\pm 0.001}$	$0.497_{\pm 0.001}$	$0.848_{\pm 0.008}$	$0.590_{\pm 0.006}$

intra modal similarity, and our proposed priors `MedCertAIn I` (inter-modal + corruptions) and `MedCertAIn II` (inter-intra + corruptions). `MedCertAIn I` corresponds to the main `MedCertAIn` variant reported in this study, as it achieves the strongest overall performance.

Gaussian priors achieve AUROC 0.820±0.004 and selective AUPRC 0.357±0.058, while random corruptions improve both standard and selective performance (AUROC 0.832±0.001, selective AUPRC 0.579±0.011). Similarity-based selection alone yields comparable standard performance (AUROC 0.821±0.001) but lower selective metrics. In contrast, `MedCertAIn I` and `MedCertAIn II` outperform all alternatives across metrics, with `MedCertAIn I` achieving the best results (AUROC 0.835±0.001, AUPRC 0.498±0.002, selective AUROC 0.857±0.005, selective AUPRC 0.599±0.002). A comparison of selective prediction trends across ablation is shown in Figure B.1. Overall, these results show that combining self-supervised latent-space divergence with corruptions produces the most informative priors and the strongest selective reliability for clinical prediction.

5.3 Impact of Multimodal Fusion on Selective Prediction

To quantify the benefit of fusion, Table 4 compares unimodal (CXR-only, EHR-only) and multimodal performance for the deterministic MedFuse baseline and our stochastic `MedCertAIn` framework. For `MedCertAIn`, unimodal variants use single-modality Random Corruptionncorruption context sets, excluding latent-space sampling for a fair comparison with MedFuse.

Across both frameworks, EHR provides the strongest unimodal signal (MedFuse: AUROC 0.760, selective AUROC 0.670; `MedCertAIn`: AUROC 0.825, selective AUROC 0.807), while CXR-only performance remains lower (MedFuse: AUROC 0.615; `MedCertAIn`: AUROC 0.632). Combining both modalities yields consistent gains: MedFuse improves to AUROC 0.819 and AUPRC 0.466, while `MedCertAIn` further increases to AUROC 0.832 and AUPRC 0.494. The largest improvements appear in selective prediction, where multimodal `MedCertAIn` achieves selective AUROC 0.841 and selective AUPRC 0.579, outperforming both EHR-only

Table 4: **Impact of Multimodal Fusion on Selective Prediction**: Quantitative performance across unimodal and multimodal settings, with consistent best performance across settings obtained with `MedCertAIn`.

Modality	Model	AUROC	AUPRC	Selective AUROC	Selective AUPRC
Unimodal CXR	MedFuse	0.615$_{\pm 0.007}$	0.225$_{\pm 0.007}$	**0.569**$_{\pm 0.012}$	**0.167**$_{\pm 0.015}$
	`MedCertAIn`	**0.632**$_{\pm 0.014}$	**0.242**$_{\pm 0.014}$	0.567$_{\pm 0.015}$	0.146$_{\pm 0.004}$
Unimodal EHR	MedFuse	0.760$_{\pm 0.003}$	0.393$_{\pm 0.008}$	0.670$_{\pm 0.017}$	0.235$_{\pm 0.007}$
	`MedCertAIn`	**0.825**$_{\pm 0.004}$	**0.480**$_{\pm 0.004}$	**0.807**$_{\pm 0.014}$	**0.499**$_{\pm 0.023}$
Multimodal	MedFuse	0.819$_{\pm 0.000}$	0.466$_{\pm 0.002}$	0.659$_{\pm 0.005}$	0.203$_{\pm 0.006}$
	`MedCertAIn`	**0.832**$_{\pm 0.001}$	**0.494**$_{\pm 0.002}$	**0.841**$_{\pm 0.006}$	**0.579**$_{\pm 0.011}$

Table 5: **Analysis of `MedCertAIn` Across Patient Subpopulations.** `MedCertAIn` consistently performs better for each individual subgroup, particularly showing significant improvements in selective AUROC and selective AUPRC, demonstrating it's potential for deployment in varied clinical scenarios and high-risk populations.

Patient Subgroup	MedFuse				MedCertAIn			
	AUROC	AUPRC	Selective AUROC	Selective AUPRC	AUROC	AUPRC	Selective AUROC	Selective AUPRC
Age								
18-45 years	0.803$_{\pm 0.001}$	**0.603**$_{\pm 0.023}$	0.559$_{\pm 0.002}$	0.262$_{\pm 0.001}$	**0.824**$_{\pm 0.001}$	0.600$_{\pm 0.002}$	**0.723**$_{\pm 0.005}$	**0.387**$_{\pm 0.017}$
45-60 years	0.788$_{\pm 0.001}$	**0.545**$_{\pm 0.027}$	0.554$_{\pm 0.003}$	0.231$_{\pm 0.001}$	**0.811**$_{\pm 0.002}$	0.485$_{\pm 0.002}$	**0.730**$_{\pm 0.008}$	**0.299**$_{\pm 0.014}$
>60 years	0.814$_{\pm 0.000}$	**0.528**$_{\pm 0.032}$	0.631$_{\pm 0.004}$	0.207$_{\pm 0.002}$	**0.832**$_{\pm 0.000}$	0.480$_{\pm 0.002}$	**0.863**$_{\pm 0.004}$	**0.559**$_{\pm 0.002}$
Sex								
Male	0.792$_{\pm 0.000}$	0.401$_{\pm 0.001}$	0.603$_{\pm 0.003}$	0.186$_{\pm 0.001}$	**0.810**$_{\pm 0.001}$	**0.425**$_{\pm 0.002}$	**0.832**$_{\pm 0.005}$	**0.494**$_{\pm 0.005}$
Female	0.806$_{\pm 0.000}$	0.420$_{\pm 0.001}$	0.609$_{\pm 0.003}$	0.183$_{\pm 0.000}$	**0.827**$_{\pm 0.001}$	**0.453**$_{\pm 0.002}$	**0.842**$_{\pm 0.005}$	**0.484**$_{\pm 0.005}$

(0.807/0.499) and CXR-only (0.567/0.146) variants. Overall, multimodal fusion improves both discrimination and selective reliability, supporting uncertainty-aware deferral decisions across unimodal and multimodal settings.

5.4 Analysis of `MedCertAIn` Across Patient Subpopulations

To evaluate robustness, we analyse performance across patient subgroups defined by age and sex. Table 5 shows that `MedCertAIn` consistently outperforms the deterministic MedFuse baseline across all subgroups, with the largest gains in selective metrics. For age groups, `MedCertAIn` improves AUROC from 0.803 to 0.824 (18–45), 0.788 to 0.811 (45–60), and 0.814 to 0.832 (>60), while selective AUROC increases from 0.559 to 0.723, 0.554 to 0.730, and 0.631 to 0.863, respectively. Although standard AUPRC is slightly lower in some age groups, selective AUPRC is consistently higher, improving from 0.262 to 0.387 (18–45), 0.231 to 0.299 (45–60), and 0.207 to 0.559 (>60).

Across sex subgroups, `MedCertAIn` also improves performance for both male and female patients. For males, AUROC increases from 0.792 to 0.810 and selective AUROC from 0.603 to 0.832, with selective AUPRC improving from 0.186 to 0.494. For females, AUROC increases from 0.806 to 0.827 and selective AUROC from 0.609 to 0.842, with selective AUPRC improving from 0.183 to 0.484. A visual comparison is provided in Figure 4.

Overall, these results indicate that `MedCertAIn` maintains strong standard performance while providing substantially more reliable selective behavior, supporting its use in high-risk populations and clinically realistic deferral settings.

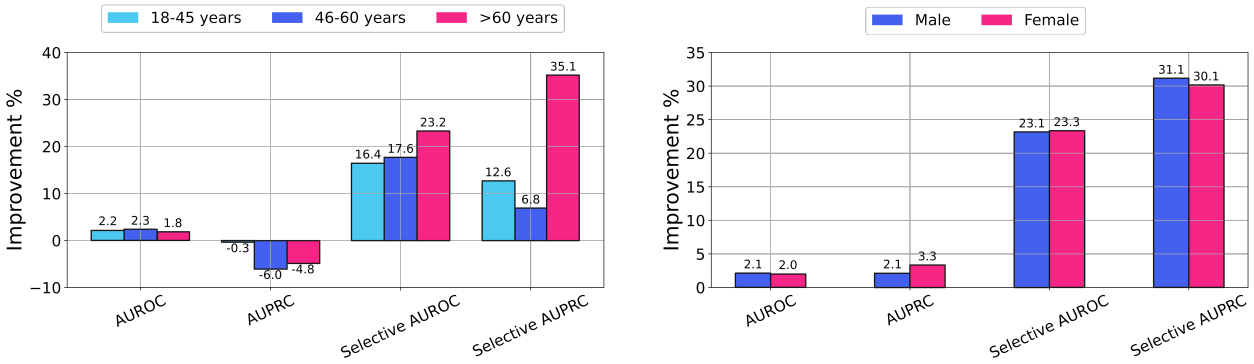

Figure 4: **Analysis of `MedCertAIn` Across Patient Subpopulations.** Comparison of percentage improvement from `MedCertAIn` over the deterministic baseline, MedFuse. We observe that our stochastic framework significantly improves selective prediction showing its adaptability over different patient csubpopulation groups.

6 Discussion

Recent studies show that UQ methods in healthcare largely rely on Bayesian approaches but are often limited to unimodal imaging settings (Seoni et al., 2023; Lechuga et al., 2025). Our work addresses this gap by applying Bayesian UQ to multimodal clinical data combining medical images and time-series signals. A key strength of `MedCertAIn` is that context point selection is performed in an unsupervised manner (i.e., without labels), making it scalable across tasks and modalities while reducing design bias. Across experiments and subpopulation analyses, `MedCertAIn` consistently improves predictive performance and selective prediction metrics over deterministic multimodal baselines, yielding more reliable uncertainty estimates.

Despite these promising results, several limitations remain. First, our experiments focus on in-hospital mortality prediction, and generalization to other clinical tasks requires further study. Second, while we integrate imaging and time-series data, incorporating additional modalities could improve applicability and performance (Salvi et al., 2023). Our corruption strategies for chest X-ray and EHR data were adapted from prior work (Rudner et al., 2022b; 2023a). While appropriate for establishing a proof of concept, future work should investigate clinically informed perturbations that better reflect real-world data quality issues and operational variability. Finally, reliance on the MIMIC dataset may limit external validity across institutions and patient populations, motivating evaluation in more diverse clinical settings (Meng et al., 2022).

Beyond performance gains, our findings suggest that uncertainty-aware multimodal models may offer practical benefits for clinical decision-making. Improved selective prediction performance indicates the potential to defer uncertain cases, which is particularly important in high-risk ICU settings. Consistent improvements across subpopulations further suggest that Bayesian approaches may contribute to more reliable model behavior across patient groups, though this warrants deeper fairness-focused investigation. Building on these results, future work should explore additional modalities such as radiology reports, extend evaluation to other clinical tasks (e.g., decompensation or length of stay), and investigate alternative fusion strategies including late fusion and missing-modality settings. Strengthening the multimodal backbone architecture and exploring self-supervised approaches for generating high-uncertainty samples may further improve robustness and uncertainty estimation.

7 Conclusion

We introduced `MedCertAIn`, a Bayesian uncertainty quantification framework for multimodal in-hospital mortality prediction that combines chest X-ray imaging and EHR time-series data. Across experiments, our approach consistently improves predictive performance and uncertainty quality compared to deterministic multimodal baselines, enabling more reliable selective prediction. These results demonstrate the potential of Bayesian multimodal learning to provide calibrated uncertainty estimates in complex clinical settings. Overall, `MedCertAIn` represents a step toward more trustworthy multimodal AI systems capable of supporting safe clinical decision-making in ICU environments.

References

Aya Awad, Mohamed Bader-El-Den, James McNicholas, and Jim Briggs. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International journal of medical informatics*, 108:185–195, 2017.

Mohammad Azizmalayeri, Ameen Abu-Hanna, and Giovanni Cinà. Unmasking the chameleons: A benchmark for out-of-distribution detection in medical tabular data. *International Journal of Medical Informatics*, 195:105762, 2025. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2024.105762. URL https://www.sciencedirect.com/science/article/pii/S1386505624004258.

Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W. Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks. 2021.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622, Lille, France, July 2015. PMLR.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. 2018.

Vince D Calhoun and Jing Sui. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(3):230–244, 2016.

Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning – a statisticians' view, 2023.

Paul KJ Han, William MP Klein, and Neeraj K Arora. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838, 2011.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, August 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013. ISSN 1532-4435.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020.

Zhicheng Jiao, Ji Whae Choi, Kasey Halsey, Thi My Linh Tran, Ben Hsieh, Dongcui Wang, Feyisope Eweje, Robin Wang, Ken Chang, Jing Wu, et al. Prognostication of patients with covid-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *The Lancet Digital Health*, 3 (5):e286–e294, 2021.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and S Horng IV. Mimic-iv-ed. *PhysioNet*, 2021.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Alain Jungo, Richard McKinley, Raphael Meier, Urspeter Knecht, Luis Vera, Julián Pérez-Beteta, David Molina-García, Víctor M Pérez-García, Roland Wiest, and Mauricio Reyes. Towards uncertainty-assisted brain tumor segmentation and survival prediction. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pp. 474–485. Springer, 2018.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof von Kalle, Titus Josef Brinker, et al. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8):e36427, 2022.

Leopoldo Julian Lechuga, Shaza Elsharief, Dhiyaa Al Jorf, Firas Darwish, Congbo Ma, and Farah E. Shamout. Uncertainty quantification for machine learning in healthcare: A survey. In Xuhai Orson Xu, Edward Choi, Pankhuri Singhal, Walter Gerych, Shengpu Tang, Monica Agrawal, Adarsh Subbaswamy, Elena Sizikova, Jessilyn Dunn, Roxana Daneshjou, Tasmie Sarker, Matthew McDermott, and Irene Chen (eds.), *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 862–907. PMLR, 25–27 Jun 2025. URL https://proceedings.mlr.press/v287/lopez25a.html.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Comput.*, 4(3): 448–472, May 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

David C Miller, Brenda Ng, John Eslick, Charles Tong, and Yang Chen. Advanced computational tools for optimization and uncertainty quantification of carbon capture processes. In *Computer Aided Chemical Engineering*, volume 34, pp. 202–211. Elsevier, 2014.

Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W. Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim G. J. Rudner, Faris Sbahi, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning, 2022.

Radford M Neal. Bayesian Learning for Neural Networks. 1996.

Venkat Nemani, Luca Biggio, Xun Huan, Zhen Hu, Olga Fink, Anh Tran, Yan Wang, Xiaoge Zhang, and Chao Hu. Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, 205:110796, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32*. 2019a.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019b.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems 35*, 2022a.

Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In *Proceedings of the 39th International Conference on Machine Learning*, 2022b.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023a.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Should we learn most likely functions or parameters? In *Advances in Neural Information Processing Systems 36*, 2023b.

Tim G. J. Rudner, Xiang Pan, Yucen Lily Li, Ravid Shwartz-Ziv, and Andrew Gordon Wilson. Uncertainty-aware priors for fine-tuning pre-trained vision and language models. In *ICML Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Reza Sadeghi, Tanvi Banerjee, and William Romine. Early hospital mortality prediction using vital signals. *Smart Health*, 9:265–274, 2018.

Massimo Salvi, Hui Wen Loh, Silvia Seoni, Prabal Datta Barua, Salvador García, Filippo Molinari, and U Rajendra Acharya. Multi-modality approaches for medical support systems: A systematic review of the last decade. *Information Fusion*, pp. 102134, 2023.

Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, pp. 107441, 2023.

Farah E Shamout, Yiqiu Shen, Nan Wu, Aakash Kaku, Jungkyu Park, Taro Makino, Stanislaw Jastrzkeski, Jan Witowski, Duo Wang, Ben Zhang, et al. An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *NPJ digital medicine*, 4(1):80, 2021.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016a.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016b.

Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions, 2022.

Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008. ISBN 1601981848.

Tong Xia, Jing Han, and Cecilia Mascolo. Benchmarking uncertainty quantification on biosignal classification tasks under dataset shift. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pp. 347–359. Springer, 2022.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16416–16424, 2024.

Yu-Dong Zhang, Zhengchao Dong, Shui-Hua Wang, Xiang Yu, Xujing Yao, Qinghua Zhou, Hua Hu, Min Li, Carmen Jiménez-Mesa, Javier Ramirez, et al. Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation. *Information Fusion*, 64:149–187, 2020.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A review of uncertainty estimation and its application in medical imaging. *arXiv preprint arXiv:2302.08119*, 2023.

Appendix

Table of Contents

A Extended Results **18**

A.1 Comparison with Hard Example Mining 18

B Comparison Across Informative Multimodal Data-Driven Priors **19**

C Derivation of Variational Objective **20**

C.1 A Family of Data-Driven Priors 20

C.2 Data-Driven, Uncertainty-Aware Priors for Fine-Tuning Pre-trained Models 20

C.3 Variational Inference with Uncertainty-Aware Priors 21

D Experimental Details **23**

D.1 Dataset Training Details 23

D.2 Training details of ConVIRT 23

D.3 Stochastic Hyperparameter Space 23

D.4 Technical Implementation 23

A Extended Results

A.1 Comparison with Hard Example Mining

Table A.1: **HEM Performance**: Performance comparison of deterministic and stochastic models using Hard Example Mining for fine-tuning during training.

Hard Example Mining (HEM)

Model	AUROC	AUPRC	Selective AUROC	Selective AUPRC
MedFuse (HEM (Shrivastava et al., 2016a))	0.233±0.003	0.094±0.001	0.393±0.010	0.070±0.010
MedCertAIn (HEM 20%)	**0.821**±0.002	**0.477**±0.002	**0.790**±0.011	**0.502**±0.023
Change (%)	58.75%	38.30%	39.65%	43.26%

We further consider Hard Example Mining (HEM), including online HEM (Shrivastava et al., 2016b) which prioritizes high-loss examples during training. In contrast, `MedCertAIn` uses high-loss samples to construct an uncertainty-aware prior rather than as training targets. We show that HEM-based fine-tuning degrades performance, while using HEM samples as context sets improves results but remains inferior to `MedCertAIn`.

The primary aim of HEM is to improve performance by focusing training on difficult samples, whereas our method targets uncertainty estimation and selective prediction, and as such, is not directly comparable to our approach. Nonetheless, to further elucidate the distinctions between `MedCertAIn` and HEM, we conducted controlled comparisons using the top 20% hardest training samples (ranked by loss). Incorporating the HEM subset as a context set within `MedCertAIn`'s uncertainty-aware prior, the model demonstrated a substantial performance boost (AUROC: 0.821, AUPRC: 0.477, Selective AUROC: 0.790, Selective AUPRC: 0.502). This represents relative gains of 58.75% AUROC, 38.30% AUPRC, 39.65% Selective AUROC, and 43.26% Selective AUPRC, underscoring the robustness of our selective prediction objective.

Unlike HEM, which emphasizes learning directly from difficult samples, `MedCertAIn` leverages them to inform uncertainty modeling while explicitly avoiding overfitting. This is enabled by our training objective, which incorporates two complementary terms: one that fits the model to the full training set, ensuring broad generalization, and another that shapes the uncertainty-aware prior using difficult examples, encouraging the model to remain calibrated and non-confident on ambiguous inputs.

B Comparison Across Informative Multimodal Data-Driven Priors

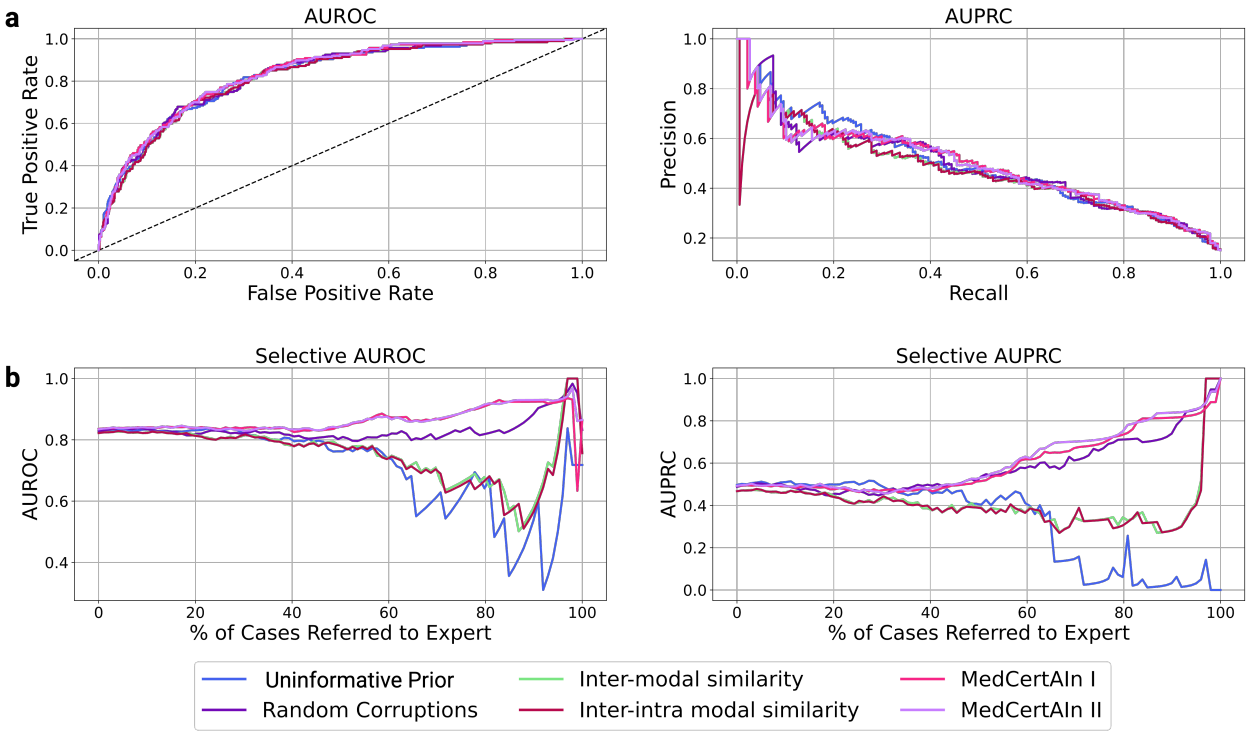

Figure B.1: **Comparison Across Informative Multimodal Data-Driven Priors. a)** Difference across baselines in standard metrics is almost negligible. **b)** The trends in selective prediction metrics show that `MedCertAIn` remains the most performing model compared to other ablations, showing that the combination of different context-set priors enhances uncertainty-ware predictions.

C Derivation of Variational Objective

We adapt the approach in Rudner et al. (2024) to the multi-modal settings. We reproduce the derivation from Rudner et al. (2024) verbatim below.

C.1 A Family of Data-Driven Priors

Consider a parametric observation model $p_{Y|X,\Theta}(y \,|\, x, \theta; f)$, and let the mapping f be defined by $f(\cdot\,;\theta) \doteq h(\cdot\,;\theta_h)\theta_L$, where $h(\cdot\,;\theta_h)$ is the post-activation output of the penultimate layer, Θ_L is the set of stochastic final-layer parameters, Θ_h is the set of stochastic non-final-layer parameters, and $\Theta \doteq \{\Theta_h, \Theta_L\}$ is the full set of stochastic parameters. We assume access to pre-trained feature parameters, θ_h^*, and context data that encodes useful information about the downstream tasks. We denote a batch of context inputs with corresponding context labels by $x_\text{c} = \{x_1, ..., x_M\}$ and $y_\text{c} = \{y_1, ..., y_M\}$, respectively, and let p_{X_c,Y_c} be a joint distribution over context batches.

To construct a family of data-driven priors, we begin by specifying a *context inference problem*. We consider a Bernoulli random variable $\check{Z}$ denoting whether a given set of neural network parameters induces predictions that exhibit some desired property (e.g., high uncertainty on a set of evaluation points). Furthermore, we define a *context observation model* $\check{p}_{\check{Z}\,|\,\Theta}(\check{z} \,|\, \theta; f, p_{X_c,Y_c})$—which denotes the likelihood of observing a yet-to-be-specified outcome $\check{z}$ under $\check{p}_{\check{Z}\,|\,\Theta}$ given θ and p_{X_c,Y_c}—and specify a *base prior* over the model parameters, $p_\Theta(\theta)$. For notational simplicity, we will drop the subscripts going forward except when needed for clarity. With this setup, we can now define the context inference problem as finding the conditional distribution over neural network parameters that we *would* obtain if we conditioned on the desired property being satisfied. This conditional distribution will serve as our data-driven prior, and by Bayes' Theorem, we can express it as

$$p(\theta \,|\, \check{z}; p_{X_c,Y_c}) = \frac{\check{p}(\check{z} \,|\, \theta; p_{X_c,Y_c}) p(\theta)}{\check{p}(\check{z}; p_{X_c,Y_c})}. \tag{C.1}$$

To define a family of data-driven priors that place high probability density on neural network parameter values that induce predictive functions with reliable uncertainty estimates, we specify a Bernoulli context observation model $\check{p}_{\check{Z}\,|\,\Theta}$ in which $\check{Z} = 1$ denotes the outcome of 'achieving reliable uncertainty quantification' and $\check{p}(\check{z} = 1 \,|\, \theta; p_{X_c,Y_c})$ denotes the likelihood of $\check{z} = 1$ given θ and p_{X_c,Y_c}. More specifically, we define:

$$\begin{aligned} \check{p}(\check{z} = 1 \,|\, \theta; p_{X_c,Y_c}) &= \exp(-\mathbb{E}_{p_{X_c,Y_c}}[c(X_c, Y_c, \theta)]) \\ \check{p}(\check{z} = 0 \,|\, \theta; p_{X_c,Y_c}) &= 1 - \check{p}(\check{z} = 1 \,|\, \theta; p_{X_c,Y_c}), \end{aligned} \tag{C.2}$$

where $c : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^P \to \mathbb{R}_{\geq}$ is a *cost function*. By specifying the outcome $\check{z} = 1$ along with a distribution, p_{X_c,Y_c} we obtain a conditional distribution $\check{p}(\theta \,|\, \check{z}; f, p_{X_c,Y_c})$—the distribution over neural network parameters that we *would* infer if we observed the outcome $\check{z} = 1$ under the base prior and the Bernoulli context observation model defined in Equation (C.2). Naturally, the quality (i.e., the usefulness) of this conditional distribution is determined by the quality of the context observation model $\check{p}_{\check{Z}\,|\,\Theta}$, the data, and the prior. As a result, the primary challenge in designing effective uncertainty-aware priors lies in constructing a context observation model—via a cost function c and a context distribution p_{X_c,Y_c}—that is as well-specified as possible. The better specified the context observation model, the more useful the data-driven prior.

C.2 Data-Driven, Uncertainty-Aware Priors for Fine-Tuning Pre-trained Models

In this section, we present a specific instantiation of an uncertainty-aware prior for fine-tuning foundation models. To define a data-driven prior $\check{p}(\theta \,|\, \check{z}; p_{X_c,Y_c})$ that incorporates useful information from the pre-trained parameters θ_h^* and assigns high probability density to parameter values θ that induce models with reliable uncertainty quantification, we need to specify a suitable context likelihood and suitable layer-specific base priors $p(\theta_h)$ and $p(\theta_L)$. For the base priors, we let $p(\theta_h) = \mathcal{N}(\theta_h; \theta_h^*, \tau_h^{-1} I)$, which assigns high probability to parameters θ_h that are close to the pre-trained parameters θ_h^*, and $p(\theta_L) = \mathcal{N}(\theta_L; \mathbf{0}, \tau_L^{-1} I)$.

To define a context observation model that induces a data-driven prior with desirable properties, we specify a cost function c of the form

$$c(x_c, y_c, \theta) \doteq \tau \sum_{k=1}^{K} D^2_{\mathcal{M}}([f(x_c;\theta)]_k, m(x_c, y_c)_k, C(x_c)), \tag{C.3}$$

where K is the number of output dimensions, p_{X_c,Y_c} is a joint distribution over context batches x_c and y_c (each of size M),

$$D^2_{\mathcal{M}}([f(x_c;\theta)]_k, m(x_c, y_c)_k, C(x_c)) \doteq \mathbf{v}_k^\top C(x_c)^{-1} \mathbf{v}_k \tag{C.4}$$

with $\mathbf{v}_k \doteq [f(x_c;\theta)]_k - m(x_c, y_c)_k$ is the squared Mahalanobis distance between model predictions $[f(x_c;\theta)]_k$ and an input-dependent distribution with mean $m(x_c, y_c)_k$ and M-by-M covariance matrix $C(x_c)$. To obtain a data-driven prior that assigns high probability density to parameters θ that induce models with reliable uncertainty estimates, we specify a data-dependent mean function, $m(x_c, y_c)_k \doteq [y_c]_k$, and a covariance function

$$C(\cdot) \doteq s_1 h(\cdot;\theta_h^*) h(\cdot;\theta_h^*)^\top + s_2 I, \tag{C.5}$$

parameterized by pre-trained model parameters θ_h^* and fixed scaling parameters τ, s_1, and s_2, that reflects structure in the pre-trained model representations $h(\cdot;\theta_h^*)$. Finally, we define the context distribution as $p(x_c, y_c) = p(y_c \mid x_c) p(x_c)$, where

$$p(y_c \mid x_c) \doteq \delta(\{\mathbf{0}, ..., \mathbf{0}\} - y_c)$$

and $p(x_c)$ is an empirical distribution constructed from a larger set of (domain- and task-specific) context inputs.[1]

Under this cost function and context distribution, the data-driven prior defined in Equation (C.2), by design, assigns high probability density to parameters θ that induce predictions $f(x_c;\theta)$ that have high predictive uncertainty on the context inputs. If the distribution over context inputs, p_{X_c}, is specified to place high probability density on context batches which contain input points that are meaningfully distinct from the training inputs, then the data-driven prior favours models that exhibit high predictive uncertainty on such meaningfully distinct inputs.

C.3 Variational Inference with Uncertainty-Aware Priors

In this section, we show how to perform variational inference with uncertainty-aware priors. We start by specifying a probabilistic model with an uncertainty-aware prior,

$$p(y_\mathcal{D}, \theta \mid x_\mathcal{D}, \check{z}; p_{X_c,Y_c}) = \underbrace{p(y_\mathcal{D} \mid x_\mathcal{D}, \theta; f)}_{\text{Likelihood}} \cdot \underbrace{p(\theta \mid \check{z}; p_{X_c,Y_c}))}_{\text{Uncertainty-aware prior}}. \tag{C.6}$$

To perform variational inference in this model and approximate the posterior distribution over the parameters of interest, we begin by defining a variational distribution,

$$q(\theta) \doteq q(\theta_h)\, q(\theta_L), \tag{C.7}$$

where $q(\theta_L) = \mathcal{N}(\theta_L; \mu_L, \Sigma_L)$ and $q(\theta_h) = \mathcal{N}(\theta_h; \mu_h, \Sigma_h)$ with learnable variational parameters $\mu \doteq \{\mu_h, \mu_L\}$ and $\Sigma \doteq \{\Sigma_h, \Sigma_L\}$, and frame the inference problem of finding the posterior $p(\theta \mid x_\mathcal{D}, y_\mathcal{D}, \check{z})$ variationally as

$$\min_{q_\Theta \in \mathcal{Q}} \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid X_\mathcal{D}, Y_\mathcal{D}, \check{Z}}), \tag{C.8}$$

where $\mathcal{Q}$ is a mean-field Gaussian variational family. This variational problem can equivalently be expressed as maximizing the variational objective

$$\bar{\mathcal{F}}(q_\Theta) \doteq \mathbb{E}_{q_\Theta}[\log p(y_\mathcal{D} \mid x_\mathcal{D}, \Theta; f)] - \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}}). \tag{C.9}$$

[1] Defining $p(y_c \mid x_c) \doteq \delta(\{\mathbf{0}, ..., \mathbf{0}\} - y_c)$ implies that under p_{X_c,Y_c}, all context batch samples have $y_c = \mathbf{0}$, and therefore, we effectively have $m(x_c, y_c)_k \doteq \mathbf{0}$ for all context batch samples.

Unfortunately, this variational objective is intractable since the data-driven prior $\check{p}(\theta \mid \check{z}; p_{X_c,Y_c})$ defined in Equation (C.1)—which is required to compute $\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}})$—is not in general tractable.

To overcome this intractability, we take advantage of the properties of the KL divergence and note that we can express $\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}})$ as:

$$\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}}) = \mathbb{E}_{q_{\Theta_h} q_{\Theta_L}} \left[\log \frac{q(\Theta_h)\, q(\Theta_L)}{p(\Theta_h)\, p(\Theta_L)} \right] - \mathbb{E}_{q_{\Theta_h} q_{\Theta_L}} [\log \check{p}(\check{z} \mid \Theta; p_{X_c,Y_c})] + \log \check{p}(\check{z}; p_{X_c,Y_c}), \tag{C.10}$$

where the intractable log-marginal likelihood $\log \check{p}(\check{z}; p_{X_c,Y_c})$ was factored out as an additive constant independent of any learnable parameters. Using this result, we can obtain a tractable lower bound

$$\mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_{\Theta \mid \check{Z}}) \geq -\mathbb{E}_{q_{\Theta_h} q_{\Theta_L}} [\log \check{p}(\check{z} \mid \Theta; p_{X_c,Y_c})] + \mathbb{D}_{\mathrm{KL}}(q_{\Theta_h} \parallel p_{\Theta_h}) + \mathbb{D}_{\mathrm{KL}}(q_{\Theta_L} \parallel p_{\Theta_L}), \tag{C.11}$$

where each KL divergence term can be computed analytically, and we can obtain an unbiased estimator of the negative log-likelihood using simple Monte Carlo estimation.

Variational Objective. Since q_{Θ_h} and q_{Θ_L} are both mean-field Gaussian distributions, we can obtain a doubly lower bounded variational objective

$$\mathcal{F}(\mu, \Sigma) \doteq \underbrace{\mathbb{E}_{q_\Theta}[\log p(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \Theta; f)]}_{\text{Expected log-likelihood}} - \underbrace{\mathbb{D}_{\mathrm{KL}}(q_{\Theta_L} \parallel p_{\Theta_L})}_{\text{Pre-training regularization}} - \underbrace{\mathbb{D}_{\mathrm{KL}}(q_{\Theta_L} \parallel p_{\Theta_L})}_{\text{Final-layer regularization}} - \underbrace{\mathbb{E}_{q_\Theta}[\mathbb{E}_{p_{X_c,Y_c}}[c(X_c, Y_c, \Theta)]]}_{\text{Uncertainty regularization}}, \tag{C.12}$$

where the cost function and context distribution are as defined above. We can estimate all expectations in the objective using simple Monte Carlo estimation, giving the final variational objective:

$$\hat{\mathcal{F}}(\mu, \Sigma) \doteq \frac{1}{J} \sum_{j=1}^{J} \log p(y_{\mathcal{D}} \mid x_{\mathcal{D}}, \theta^{(j)}; f) - \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_\Theta) - \frac{1}{JJ'} \sum\nolimits_{j=1}^{J} \sum\nolimits_{j'=1}^{J'} c(x_c^{(j')}, y_c^{(j')}, \theta^{(j)}), \tag{C.13}$$

with $\theta^{(j)} \sim q_\Theta$, $x_c^{(j')} \sim p_{X_c}$, and $y_c^{(j')} \sim p_{Y_c|X_c}$ for $j = 1, ..., J$ and $j' = 1, ..., J'$. This objective can be maximized with stochastic variational inference (Hoffman et al., 2013).

D Experimental Details

D.1 Dataset Training Details

Table D.2: **Summary of Original Datasets.** Number of unimodal and multimodal data points available in MIMIC-IV (EHR) and MIMIC-CXR (CXR) (Johnson et al., 2019; 2021) for the in-hospital mortality task. We note that the size of the multimodal dataset for our task decreases since we drop all data samples that do not have both modalities.

Dataset	Training	Validation	Testing
Unimodal CXR	124,671	8,813	20,747
Unimodal EHR	42,628	4,802	11,914
Multimodal (CXR + EHR)	4,485	488	1,242

D.2 Training details of ConVIRT

Contrastive VIsual Representation Learning from Text (ConVIRT) (Zhang et al., 2022) is a vision-language self-supervised pretraining framework developed for medical images and radiology reports. ConVIRT is based on a bidirectional contrastive objective between pretraining modalities, maximizing the similarity of the latent space embeddings of an image-text pair. The loss function of our ConVIRT-MedFuse architecture is defined as $\mathcal{L} = \mathcal{S}(x^{\text{ehr}}, x^{\text{cxr}}) + \mathcal{S}(x^{\text{cxr}}, x^{\text{ehr}})$, where $\mathcal{S}$ is the infoNCE loss.

We conducted 10 hyperparameter tuning runs via random search by sampling a learning rate from a uniform distribution in the range $[10^{-1}, 10^{-2}]$. We used a batch size of 256 across experiments, setting the maximum number of epochs to 300, with no early stopping. To select the best model, we chose the best checkpoint based on the best AUROC score achieved on the validation set across epochs and models. Using the best model, we made an inference pass using the original training data $(X^{\text{ehr}}, X^{\text{cxr}})$ to obtain the latent space representation and compute the cosine similarity between data modalities.

D.3 Stochastic Hyperparameter Space

Table D.3: **Hyperparameter Search Space.** Hyperparameters and corresponding value ranges used for the uncertainty regularizer of `MedCertAIn`.

Hyperparameter	Values
prior variance	[0.1, 1, 10, 1000]
prior likelihood scale	[0.1, 1, 10]
prior likelihood f-scale	[0, 1, 10]
prior likelihood covariance scale	[0.1, 0.01, 0.001]
prior likelihood covariance diagonal	[0.5, 1, 5]

D.4 Technical Implementation

Our data loading and pre-processing pipeline was implemented using PyTorch, (Paszke et al., 2019) following the same structure of the code used by Hayat et al. (2022). However, we refactored the original unimodal and multimodal models, training, and evaluation loops using Jax (Bradbury et al., 2018). This framework simplifies the implementation of Bayesian neural networks and stochastic training, which are the basis of the uncertainty quantification methods used in this work. In addition, thanks to this code refactoring, we obtained a significant reduction in total training time for the unimodal and multimodal models compared to the code baseline in PyTorch.

We note that due to specific caching procedures of the Jax framework, each x_{ehr} instance has to be standardized into the same time-step length for the LSTM encoder to avoid out-of-memory issues. The Jax

framework requires that an LSTM encoder defines a static length of the sequences it is going to process, and then it caches this model in order to increase the training speed. If different sequence lengths are used, then Jax caches an instance of the LSTM encoder for each specific length to be used during each training cycle. The problem arises when dealing with a dataset that contains sequences of lengths that present high variance (i.e, many different sequence lengths for every data point in the dataset) In comparison, PyTorch does not use this approach and is able to process sequences of dynamic length with one single instance of the LSTM encoder. However, this comes at the cost of increased training times when comparing both frameworks. This problem was during the experimentation phase of our work, however we note that for the patient in-hospital mortality task the length of all x_{ehr} instances is standardized to 48hrs which does not generate this issue.