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

L. Julián Lechuga López[1,2,†] **Tim G. J. Rudner**[3] **Farah E. Shamout**[1,2]

[1]New York University
[2]New York University Abu Dhabi
[3]University of Toronto
[†]Correspondence to leopoldo.julian@nyu.edu

Abstract

Safe predictions are essential for deploying machine learning models in clinical decision support systems, requiring reliable uncertainty estimation. However, current models often lack well-calibrated uncertainty, particularly in multimodal settings where effective information fusion is needed. We propose MedCertAIn, a predictive uncertainty framework for multimodal in-hospital risk prediction. Our approach introduces data-driven priors over neural network parameters using a hybrid strategy based on cross-modal similarity in self-supervised latent representations and modality-specific data corruptions. We evaluate our method on clinical time-series and chest X-ray data from MIMIC-IV and MIMIC-CXR. MedCertAIn improves both predictive performance and uncertainty estimation compared to deterministic and Bayesian baselines. These results highlight the potential of data-driven priors for reliable, uncertainty-aware clinical AI.

1. Introduction

Trustworthy deployment of machine learning models in healthcare requires reliable uncertainty estimation due to the safety-critical nature of clinical decision-making (Begoli et al., 2019; Gruber et al., 2023; Nemani et al., 2023). Uncertainty arises from multiple sources, including model parameters, data noise, and distribution shifts, all of which are inherent in real-world clinical settings (Miller et al., 2014). Accurate uncertainty estimates can improve both model reliability and clinician trust (Seoni et al., 2023).

However, uncertainty quantification (UQ) in healthcare remains limited and fragmented (Kompa et al., 2021; Lechuga et al., 2025; Han et al., 2011). Existing approaches face challenges such as limited theoretical guidance, difficulty scaling to real-time systems, and the lack of well-constructed priors, particularly under dataset shift (Begoli et al., 2019; Zou et al., 2023; Ovadia et al., 2019). In addition, clinical decision-making is inherently multimodal, requiring models to integrate heterogeneous data sources (Saab et al., 2024). Despite this, most UQ work focuses on unimodal settings, especially medical imaging, leaving uncertainty estimation in multimodal clinical tasks largely unexplored (Gawlikowski et al., 2021; Tran et al., 2022).

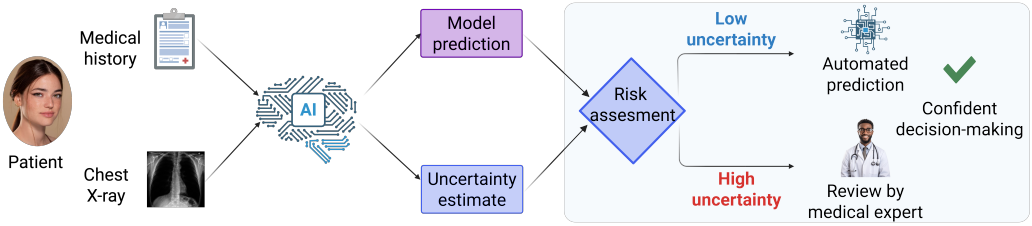

Figure 1: Uncertainty-Aware Clinical Decision-making. Patient multimodal data is processed by our uncertainty-aware model to produce a prediction and an uncertainty estimate. Clinicians accept confident predictions or defer uncertain cases for further review, improving workflow efficiency and supporting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

To address these challenges, we introduce MedCertAIn, a multimodal uncertainty-aware framework for deterioration risk prediction that enables selective prediction by deferring low-confidence cases for human review (Figure 1). Our approach combines Bayesian learning with automated, label-free priors, enabling scalable uncertainty estimation without expert annotation.

We summarize our key contributions as follows:

1. We propose MedCertAIn, a multimodal uncertainty-aware framework based on Bayesian learning and variational inference for improved reliability.

2. We introduce a label-free pipeline to construct data-driven multimodal priors using cross-modal similarity in self-supervised latent representations and modality-specific perturbations.

3. We evaluate our method on multimodal clinical data from MIMIC-IV (Johnson et al., 2021) and MIMIC-CXR (Johnson et al., 2019), showing improved predictive performance and uncertainty estimation.

4. We implement our framework in JAX and release code to support future research: https://anonymous.4open.science/r/medcertain_tmlr-8154.

2. Related Work

Multimodal learning has been widely used in healthcare to combine heterogeneous data sources such as imaging and clinical time-series, improving predictive performance in tasks like risk prediction (Hayat et al., 2022; Muhammad et al., 2021). However, multimodal integration introduces additional sources of uncertainty due to modality heterogeneity and misalignment (Azizmalayeri et al., 2025; Kurz et al., 2022). Uncertainty quantification (UQ) has been extensively studied in medical imaging (Gawlikowski et al., 2021; Seoni et al., 2023), but remains largely limited to unimodal settings. Existing methods often fail to jointly address dataset shift, model uncertainty, and multimodal inconsistencies (Lechuga et al., 2025; Liang et al., 2023). Our work builds on Bayesian neural networks and variational inference (MacKay, 1992; Neal, 1996), extending them to multimodal settings with data-driven priors for improved uncertainty estimation.

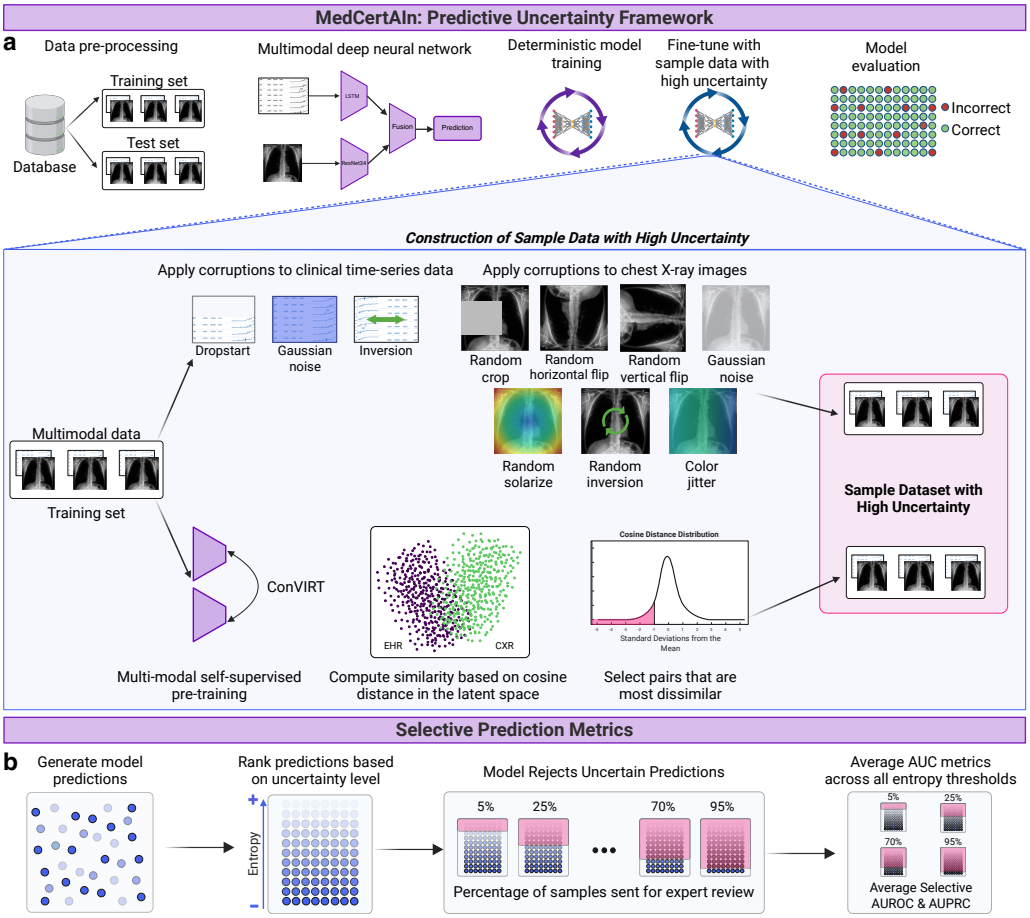

Figure 2: MedCertAIn: Predictive uncertainty framework. a) Multimodal training with data-driven priors constructed from corrupted samples and low cross-modal similarity pairs, followed by Bayesian fine-tuning. **b)** Selective prediction using entropy to defer uncertain cases, evaluated via selective AUROC/AUPRC.

3. Methods

MedCertAIn is a multimodal uncertainty-aware framework designed to produce calibrated predictions and defer low-confidence cases. Rather than modifying model architectures, our approach introduces data-driven priors and an uncertainty-aware training objective, enabling reliable selective prediction across multimodal models.

3.1. Learning Objective

We consider a supervised multimodal dataset $\mathcal{D} = \{(x_n^{ehr}, x_n^{cxr}, y_n)\}_{n=1}^N$, where EHR time-series and chest X-ray images are encoded by modality-specific encoders and fused for prediction. To improve uncertainty estimation, we construct data-driven priors conditioned on a context set X_c representing high-uncertainty samples. These samples capture distributional shifts and ambiguous inputs, encouraging the model to remain uncertain in such regions.

We optimize a variational objective augmented with an uncertainty regularization term:

$$\mathcal{F}(q_\Theta) = \mathbb{E}_{q_\Theta}[\log p(y_\mathcal{D} \mid x_\mathcal{D}, \Theta)] - \mathbb{D}_{\mathrm{KL}}(q_\Theta \parallel p_\Theta) + \mathbb{E}_{q_\Theta}\mathbb{E}_{p_{X_c}}[\log \tilde{p}(\mathbf{0} \mid X_c, \Theta)].$$

This objective encourages accurate predictions on training data while promoting high uncertainty on context samples. Optimization is performed using Monte Carlo estimation and standard variational inference techniques. Additional derivations of the tractable objective under the proposed context-conditioned prior are provided in Appendix B.

3.2. Context Set Construction

We construct a context set representing high-uncertainty regions using two label-free strategies:

1) Data perturbations. We apply controlled corruptions to both modalities (e.g., noise, temporal distortions, and image augmentations) to simulate distribution shifts.

2) Cross-modal similarity. We select samples with low agreement between modality representations in a self-supervised latent space, assuming that modality mismatch indicates higher uncertainty.

The final context set combines corrupted samples and low-similarity pairs, enabling data-driven prior construction without expert annotation.

3.3. Uncertainty Measurement

We quantify predictive uncertainty using the entropy of the predictive distribution:

$$\mathcal{H}(p(y \mid x)) = -\sum_c p(y = c \mid x) \log p(y = c \mid x). \tag{3.1}$$

Entropy is used to rank predictions for selective prediction, allowing the model to defer uncertain cases for clinical review.

4. Experimental Setting

Dataset. We evaluate on multimodal data from MIMIC-IV (EHR) and MIMIC-CXR (chest X-rays). We focus on in-hospital mortality prediction after 48 hours in the ICU, using 6,215 paired samples with an 80/20 train-test split and cross-validation on the training set.

Model. We use MedFuse (Hayat et al., 2022) as the multimodal backbone, combining an LSTM encoder for EHR and a ResNet-34 for images, followed by early fusion.

Table 1: MedCertAIn **Achieves State-Of-The-Art Performance and Reliable Selective Prediction**: Quantitative performance of the deterministic and naive stochastic baselines compared to MedCertAIn.

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., 2022)	$0.820_{\pm 0.004}$	$0.485_{\pm 0.006}$	$0.711_{\pm 0.034}$	$0.357_{\pm 0.058}$
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%

Baselines. We compare against MedFuse (deterministic), DrFuse (Yao et al., 2024), and a Bayesian neural network with an uninformative gaussian prior. All models are evaluated under identical settings.

Evaluation. We report AUROC and AUPRC, along with selective prediction metrics. Predictions are ranked by entropy, and performance is evaluated across rejection thresholds, reporting selective AUROC and AUPRC.

5. Results

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

Table 1 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., 2022), 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 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 A.1 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. Extended results comparing performance across different multimodal priors, data modalities and patient subpopulations are provided in Appendix A.

6. Discussion

Uncertainty quantification in healthcare has largely focused on unimodal settings, particularly medical imaging (Seoni et al., 2023; Lechuga et al., 2025). We address this gap with a multimodal Bayesian framework using data-driven priors for improved uncertainty estimation. A key advantage of `MedCertAIn` is its label-free construction of high-uncertainty context points, enabling scalable and less biased prior design. Our results show consistent improvements in predictive performance and selective prediction.

Several limitations remain. We focus on in-hospital mortality prediction, and extension to other tasks is needed. Incorporating additional modalities and evaluating beyond MIMIC are important directions for improving generalization (Meng et al., 2022).

Our findings highlight the practical value of uncertainty-aware multimodal models for clinical decision-making. Improved selective prediction enables deferral of uncertain cases, supporting safer deployment. Future work should explore additional modalities, fusion strategies, and improved context construction.

References

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.

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*, pages 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.

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.

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*, pages 479–503. PMLR, August 2022.

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.

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.

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

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, editors, *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 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, pages 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.

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, 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, 2019.

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.

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*, 2022.

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.

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*, page 107441, 2023.

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*, pages 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*, pages 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.

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, pages 16416–16424, 2024.

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*, pages 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 **10**

A.1 Comparison Across Informative Multimodal Data-Driven Priors 10

A.2 Impact of Multimodal Fusion on Selective Prediction 12

A.3 Analysis of MedCertAIn Across Patient Subpopulations 13

A.4 Comparison with Hard Example Mining 13

B Derivation of Variational Objective **15**

B.1 A Family of Data-Driven Priors 15

B.2 Data-Driven, Uncertainty-Aware Priors for Fine-Tuning Pre-trained Models 16

B.3 Variational Inference with Uncertainty-Aware Priors 17

C Experimental Details **19**

C.1 Dataset Training Details 19

C.2 Training details of ConVIRT 19

C.3 Stochastic Hyperparameter Space 19

C.4 Technical Implementation 19

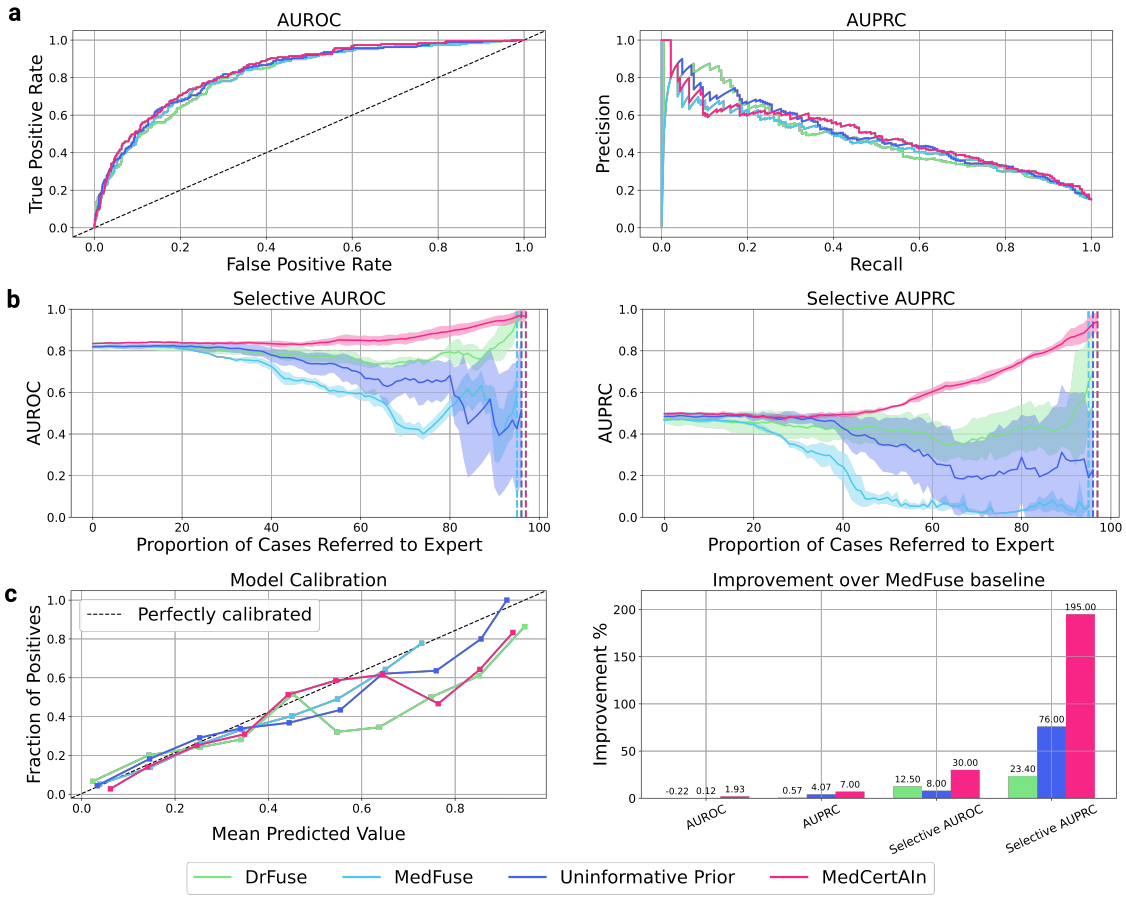

Figure A.1: MedCertAIn Achieves State-Of-The-Art Performance and Reliable Selective Prediction. **a)** In standard metrics, the differences across baselines are very small for predictive performance. **b)** Analyzing the selective prediction metrics trends, we observe larger variance for each baseline. 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.

A. Extended Results

A.1. 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 A.1 compares a naive Gaussian prior, random corruptions, inter-modal similarity, inter-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.

Table A.1: 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	$\mathbf{0.835}_{\pm \mathbf{0.001}}$	$\mathbf{0.498}_{\pm \mathbf{0.002}}$	$\mathbf{0.857}_{\pm \mathbf{0.005}}$	$\mathbf{0.599}_{\pm \mathbf{0.002}}$
MedCertAIn II	$0.834_{\pm 0.001}$	$0.497_{\pm 0.001}$	$0.848_{\pm 0.008}$	$0.590_{\pm 0.006}$

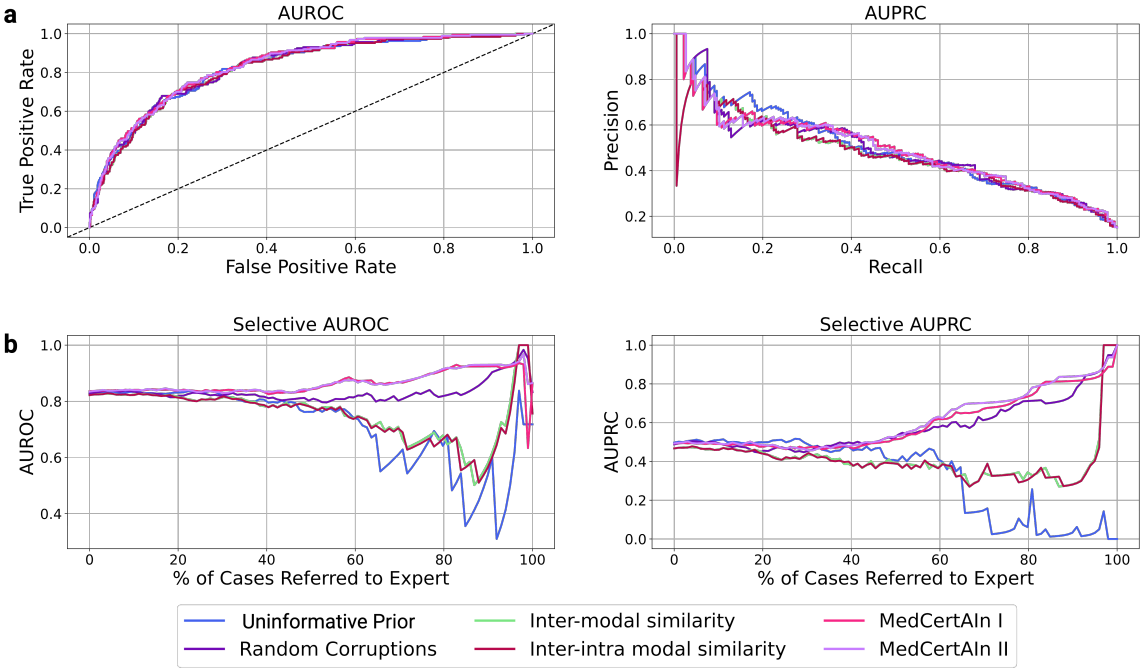

Figure A.2: Comparison Across Informative Multimodal Data-Driven Priors. a) Difference across baselines in standard metrics is almost negligible. **b)** The trends in selective prediction metrics show that MedCertAIn remains the most performing model compared to other ablations, showing that the combination of different context-set priors enhances uncertainty-ware predictions.

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

Table A.2: 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±0.007	0.225±0.007	**0.569±0.012**	**0.167±0.015**
	MedCertAIn	**0.632±0.014**	**0.242±0.014**	0.567±0.015	0.146±0.004
Unimodal EHR	MedFuse	0.760±0.003	0.393±0.008	0.670±0.017	0.235±0.007
	MedCertAIn	**0.825±0.004**	**0.480±0.004**	**0.807±0.014**	**0.499±0.023**
Multimodal	MedFuse	0.819±0.000	0.466±0.002	0.659±0.005	0.203±0.006
	MedCertAIn	**0.832±0.001**	**0.494±0.002**	**0.841±0.006**	**0.579±0.011**

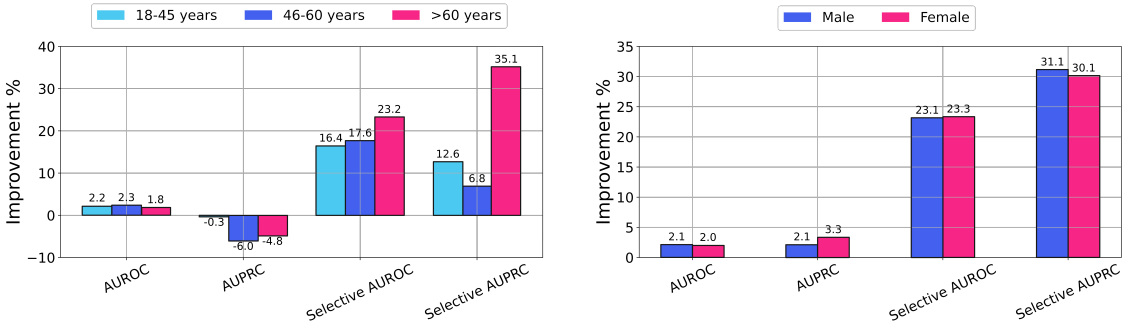

Figure A.3: Analysis of MedCertAIn Across Patient Subpopulations. Percentage improvement from MedCertAIn over the deterministic baseline, MedFuse. Our stochastic framework significantly improves selective prediction showing its adaptability over different patient subpopulation groups.

of selective prediction trends across ablation is shown in Figure A.2. 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.

A.2. Impact of Multimodal Fusion on Selective Prediction

To quantify the benefit of fusion, Table A.2 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 (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.

Table A.3: 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**$_{\pm 0.005}$

A.3. Analysis of MedCertAIn Across Patient Subpopulations

To evaluate robustness, we analyse performance across patient subgroups defined by age and sex. Table A.3 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 A.3.

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.

A.4. Comparison with Hard Example Mining

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,

Table A.4: 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))	$0.233_{\pm 0.003}$	$0.094_{\pm 0.001}$	$0.393_{\pm 0.010}$	$0.070_{\pm 0.010}$
MedCertAIn (HEM 20%)	$\mathbf{0.821}_{\pm 0.002}$	$\mathbf{0.477}_{\pm 0.002}$	$\mathbf{0.790}_{\pm 0.011}$	$\mathbf{0.502}_{\pm 0.023}$
Change (%)	58.75%	38.30%	39.65%	43.26%

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. 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.

B.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

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

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:

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

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 (B.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.

B.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 $\check{p}(\theta \mid \check{z}; p_{X_c,Y_c})$ 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

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

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),

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

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

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

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 (B.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.

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.

B.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,

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

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

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

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

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

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

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

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 (B.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:

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

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

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

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

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

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:

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

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).

C. Experimental Details

C.1. Dataset Training Details

Table C.5: 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

C.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.

C.3. Stochastic Hyperparameter Space

Table C.6: 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]

C.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.