MedCertAIn: Uncertainty-Aware Multimodal AI for Trustworthy In-Hospital Mortality Prediction

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

[1] New York University, [2] New York University Abu Dhabi

Main Takeaways

- Trustworthy and reliable AI in healthcare requires the ability to **communicate uncertainty** for individual patient predictions.
- `MedCertAIn` **integrates multiple modalities** without sacrificing predictive performance through principled uncertainty estimation for in-hospital mortality prediction.
- Our framework can help **optimize clinical workflows**, by allowing clinicians to spend more time on more complex cases.
- The empirical findings show the potential of `MedCertAIn` for **safe deployment** of AI-based clinical decision support systems, by enhancing transparency and fostering trust.

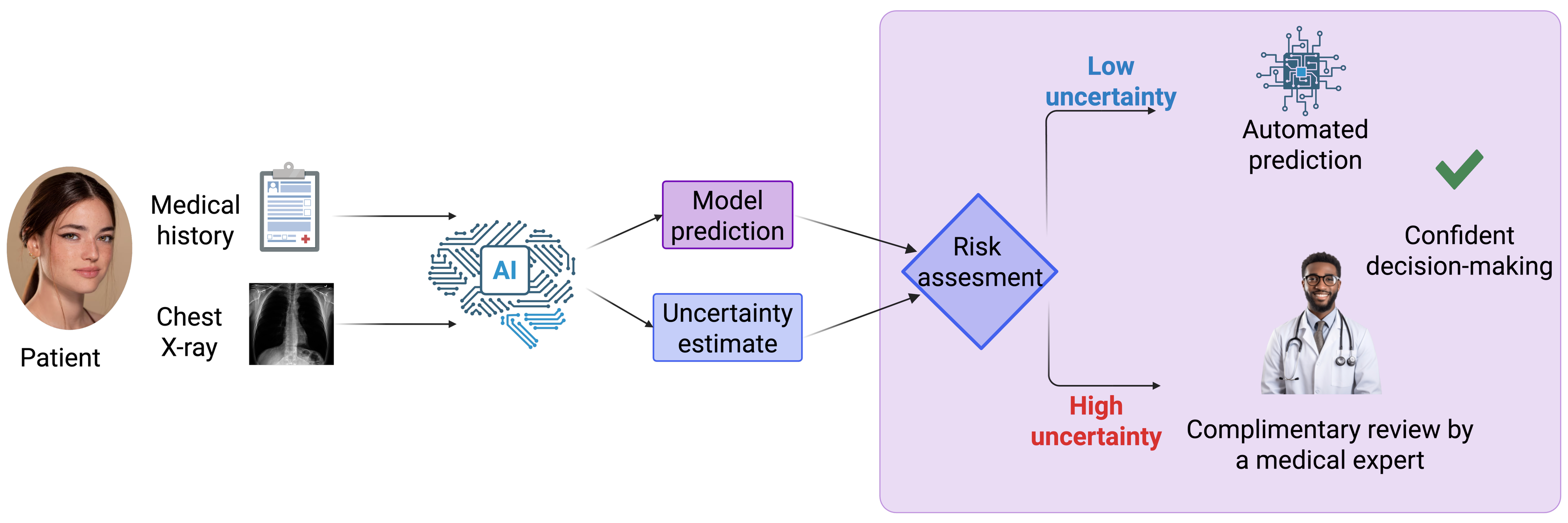


Figure 1: Uncertainty-Aware Clinician-AI Cooperation Framework.

Technical Details

- **Bayesian Learning Framework:** `MedCertAIn` enhances uncertainty quantification for multimodal in-hospital mortality prediction.
- **Real-World Clinical Data:** MIMIC-IV (i.e time-series data from labs, vitals, etc) and MIMIC-CXR (i.e. chest x-rays), integrating structured clinical datasets encouraging the model to learn complementary patterns across modalities.
- **Multimodal Architecture:** A two-layer LSTM network and a ResNet-34, combined using a Fusion layer.
- **Empirical Bayes Variational Inference:** We optimize the model's predictive distribution while regularizing it to remain cautious on unfamiliar inputs.
- **Robustness to Data Shifts:** `MedCertAIn` simulates out-of-distribution samples to train the model to detect unusual and low-confidence medical cases.

Learning to Detect Difficult Cases

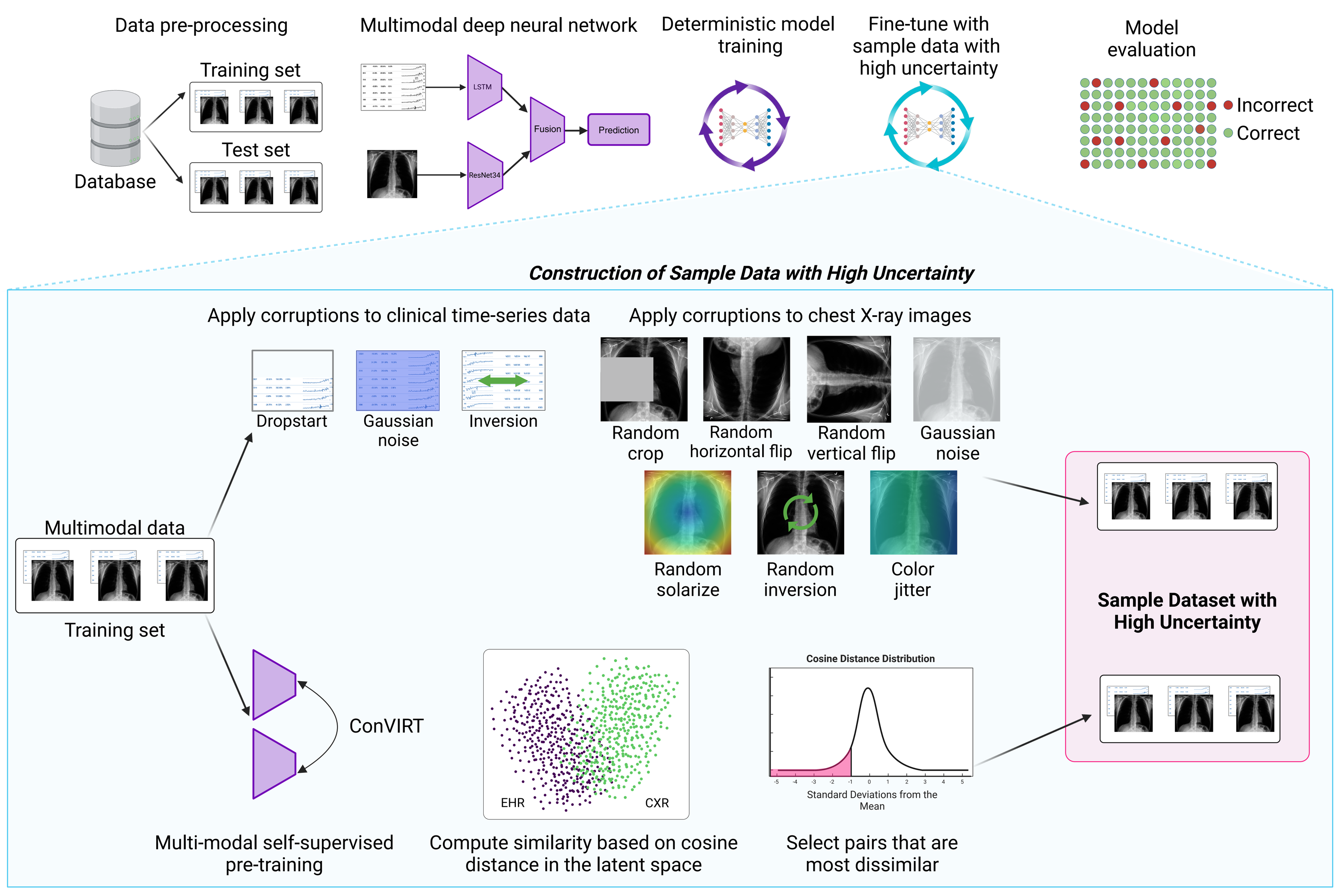


Figure 2: Model Fine-tuning on High-Uncertainty Patients.

- We create "challenging" samples by adding controlled noise and identifying cases where patient modality information seems inconsistent with each other.
- These high-uncertainty cases help the model learn to make predictions and recognize when it might less confident about a patient.
- This improves performance and ensures clinicians are alerted when closer review of a patient may be needed.

Evaluating Model Confidence

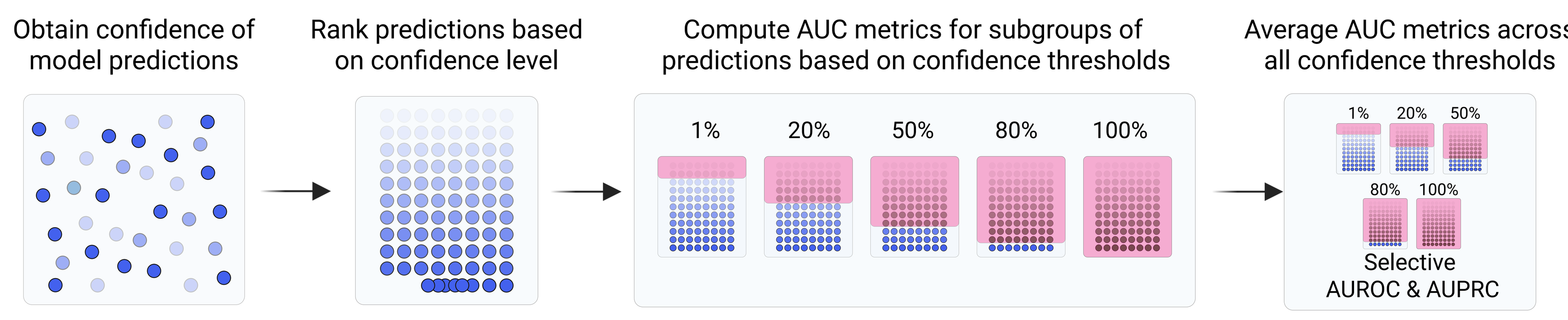
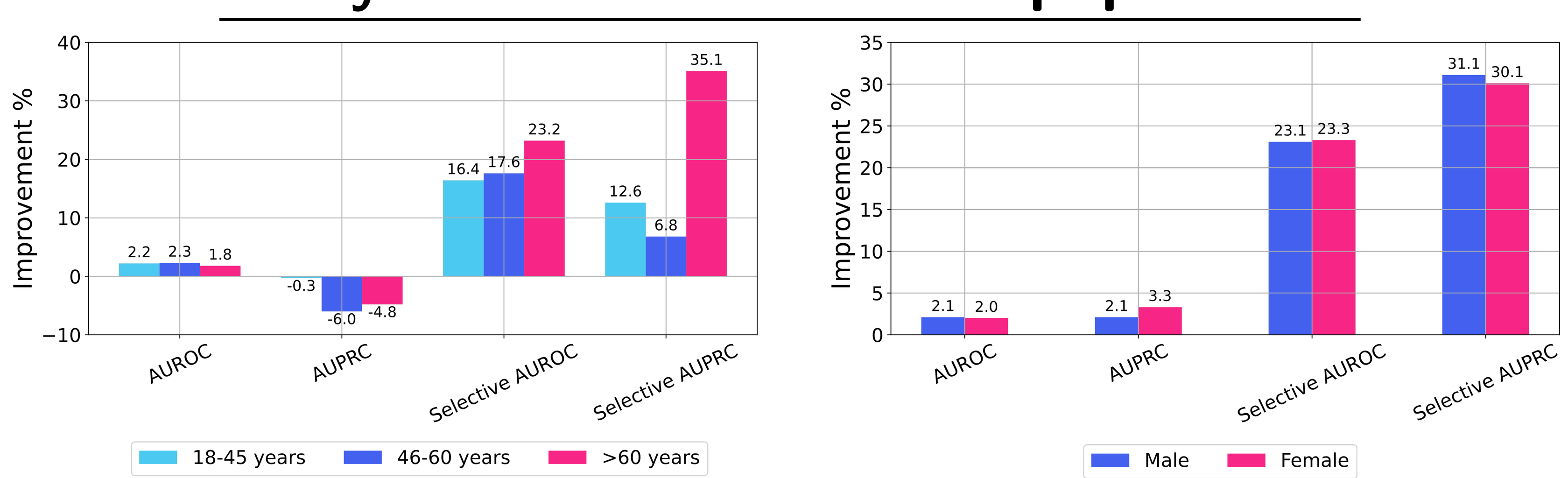


Figure 3: Selective Prediction Metrics.

- Selective prediction measures the model's ability to detect low-confidence predictions.
- A varying rejection threshold of "difficult" patients calibrates when a clinician should step in.
- This evaluation strengthens clinical applicability and trust on uncertainty-aware models.

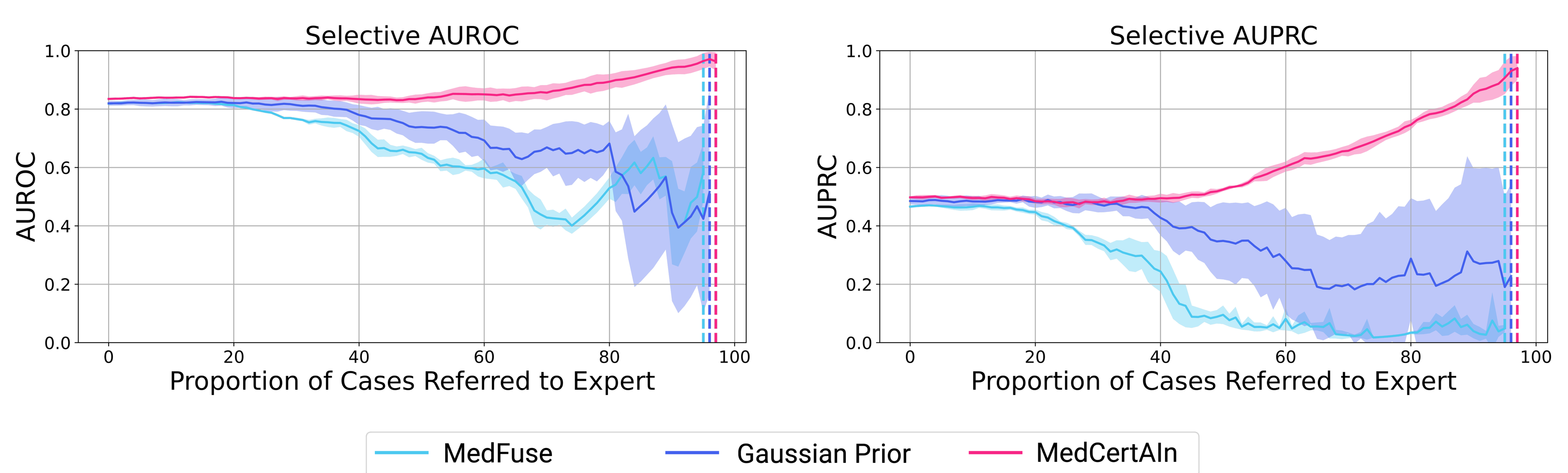
Respiratory Shock Detection Performance

Analysis Across Patient Subpopulations



- Compared to our baseline MedFuse, `MedCertAIn` shows consistent improvement in selective AUROC and selective AUPRC, for each individual subpopulation.
- This demonstrates `MedCertAIn`'s potential for deployment in varied clinical scenarios and high-risk populations.

Improving Reliability of Clinician-AI Cooperation



- Compared to the baseline MedFuse, `MedCertAIn` presents a significant improvement of 30.0% and 195% in selective AUROC and selective AUPRC, respectively.
- This demonstrates `MedCertAIn`'s enhanced ability to detect patients for which it struggles to obtain a correct prediction.