ICU Mortality Prediction

Qing Shen [qshen19], Léonard Mauvernay [lmauvern], Antong Zhang [azhan307]

Introduction

ICU datasets can look very different from one hospital to another, especially in how often measurements are recorded and how variables are defined. Because of this, a model that performs well in one ICU setting may not work as well in another. We want to explore this problem using modern deep learning methods, so our project focuses on Transformer-based time-series models and how well they transfer across datasets.

We train our main model on the HiRID dataset, which contains high-resolution ICU data that tends to work well with deep learning architectures. After building a strong baseline on HiRID, we test how well the same model performs on MIMIC-IV, a widely used ICU dataset with different recording patterns and clinical characteristics. We also try several domain-adaptation strategies to see whether we can narrow the performance gap between training on one dataset and applying the model to another.

Overall, this is a supervised learning problem and a binary classification task. Using the first 24 hours of ICU data, we predict whether a patient will die during their ICU stay.

Related Works

One relevant article that we found was entitled: Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises. The review provides a clear overview of how researchers use datasets such as MIMIC-IV and identifies what kind of models are normally used for ICU predictions. According to the authors, many studies apply several deep learning methods, including GRUs, LSTMs, and Transformer-style architectures, to address the irregular and time-varying nature of ICU data. The review mentions other more traditional techniques like logistic regression and random forests, which are more often used as baselines (Khaled, 2025).

Datasets

We use two large, de-identified ICU EHR datasets from PhysioNet: MIMIC-IV v3.1 (Beth Israel Deaconess Medical Center; ~365,000 patients, ~546,000 hospitalizations, ~94,000 ICU stays) and HiRID v1.1.1 (Bern University Hospital; ~34,000 ICU admissions with 681 high-frequency variables). Access to both datasets is restricted and requires credentialing and a Data Use Agreement. No new data will be collected. Extensive preprocessing will be carried out, including the following: (a) restricting to adult ICU stays with adequate observation time (b) converting each stay into fixed-length time-series windows on a common time grid (c) handling missing values and implausible measurements and (d) harmonizing variable definitions and units across MIMIC-IV and HiRID in order to enable cross-dataset analysis.

Method and Metrics

We will build a Transformer based time-series model for ICU mortality prediction. We will use a following the Transformer model as our backbone and then extending it with domain adaptation. For each patient stay, we construct a multivariate time series over the first 24 hours after ICU admission with hourly bins; each time step contains vitals, labs, basic interventions, a missingness mask, and time-since-last-measure features. These vectors are linearly projected to a shared embedding space, augmented with positional embeddings, and passed through a stack of encoder layers . The final hidden representation is pooled and fed into an MLP classifier to output the probability of ICU mortality. We first train and tune this model on HiRID only, which is our source domain. This gives us a strong in-domain baseline that should roughly match other HiRID Transformer performance.

The novel part of the project is the domain adaptation where we intend to mimic a scenario where we are deploying our model in a hospital, which will have limited data. After pretraining on HiRID, we explore several domain adaptation methods on MIMIC IV. The most straightforward method would be simple fine-tuning of the model on a small labeled subset of MIMIC IV. We will also use domain adversarial neural network where we add a domain classifier on top of the shared encoder and train with a gradient reversal layer to encourage domain invariant representations. The hardest parts will likely be robust preprocessing/harmonization (aligning variable definitions and scaling between HiRID and MIMIC-IV) and stabilizing adversarial training. As backups, we will consider using datasets that are easier for cross-dataset comparison.

Success is primarily defined by how well the model predicts patient mortality on the target domain, especially when labeled target data are scarce, and by whether domain adaptation meaningfully narrows the performance gap to a fully supervised target model. Because this is a highly imbalanced clinical risk prediction problem, raw accuracy is not appropriate; we will instead focus on AUROC and AUPRC as primary metrics. Our experiments will include: 1) in-domain baselines; 2) cross-domain performance with no adaptation; and 3) transfer learning and domain adaptation experiments where we vary the size of the labeled MIMIC IV subset and compare target-only training, HiRID pretrain + fine-tune, and DANN variants. The base goal is to correctly reproduce competitive in-domain performance on HiRID and show some improvements after fine-tuning. The more ambitious goal would be establishing the practical method for deploying the model in hospitals.

Ethics

What broader societal issues are relevant to your chosen problem space?

ICU mortality prediction is tied to triage and how scarce resources like beds, staff, and machines are used. If such models are deployed, they can influence who gets closer monitoring or escalation of care, which raises fairness questions across age, race, and socioeconomic status.

The data come from two high-income hospitals, so models based on them may work worse in other settings. And because we rely on large EHR datasets collected without case-by-case consent, there are ongoing issues around data governance, privacy, and who benefits from the resulting models.

Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?

Stakeholders include ICU patients and families, clinicians using risk scores, and hospitals that might deploy them. A false negative could mean a high-risk patient is not flagged and care escalation is delayed. A false positive could add to alarm fatigue or lead to unnecessary tests and treatments. In our project, we only train and evaluate models offline on de-identified data, so there is no direct impact on patients. But any real-world use would need strong local validation, calibration checks, and clear framing of the model as decision support, not an automatic gatekeeper.

Division of Labor

To set up the project, Qing will be responsible for creating the credential account and completing CITI training to acquire access to the MIMIC IV and HiRID datasets, while Leo and Antong will be setting up the transformer model and familiarizing with the framework.

Moving on, we will work collaboratively on training the transformer on the source and target domains, and the finetuning/DANN model for domain adaptation.

Stephen was “traveling today” and said nothing else; hence, we have no information regarding what role he wants to play in this project.

Reference

Khaled, A., Sabir, M., Qureshi, R., Caruso, C. M., Guarrasi, V., Xiang, S., & Zhou, S. K. (2025). *Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises*. arXiv preprint arXiv:2506.12808.