

Active Learning Strategies for Imbalanced Clinical Note Categories in ICU Mortality Prediction

Mina Afzali*, Mohsen Kahani, Maryam Karrabi, Nadieh Armin

*- Corresponding author: Bachelor Student in Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

2- Professor of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

m.afzali@mail.um.ac.ir

Abstract

- ICU mortality prediction is critical for clinical decision-making.
- Clinical notes provide rich information but suffer from **severe class imbalance**.
- We analyze **14 clinical note categories** from MIMIC-III.
- A class-aware active learning framework with **GatorTron embeddings** is proposed.
- Achieves comparable or higher AUC using only 15–38% of the data.
- Nursing-related notes** show the strongest predictive contribution.

Research method

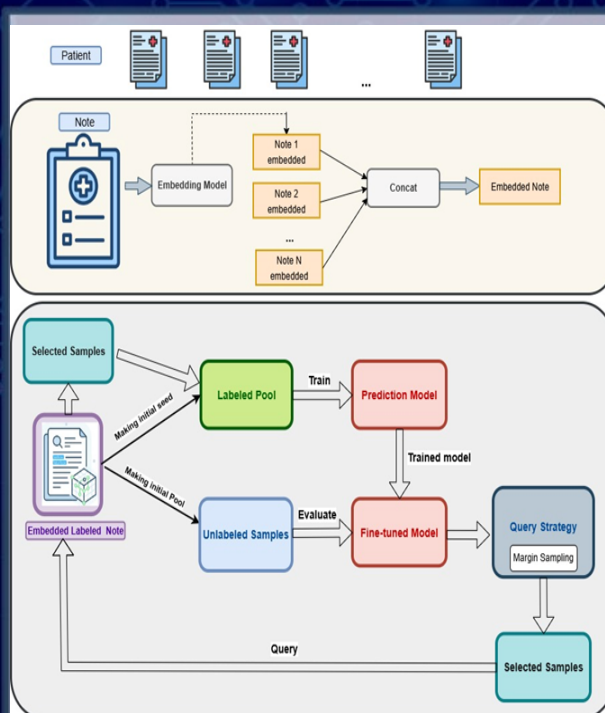


Fig.1. The proposed Method for mortality prediction.

conclusion

- Clinical note categories provide strong signals for ICU mortality prediction.
- Nursing-related notes** are the most informative single category.
- Class-aware active learning** effectively handles severe data imbalance.
- Comparable or superior AUC is achieved with substantially less data.
- The framework enables efficient and scalable clinical prediction.
- Supports practical deployment in **resource-limited healthcare settings**.

The purpose of the research

- To apply **active learning** for reducing training data usage in ICU mortality prediction
- To show that selective sampling can achieve **equal or better performance** with less data
- To mitigate **severe class imbalance** through class-aware sample selection
- To compare the predictive contribution of clinical note categories
- To develop a data-efficient framework for mortality prediction from clinical text

Practical or simulation results

TABLE I. Active vs. Passive Learning Performance

Category	Passive AUC	Active AUC	Data Used	Key Improvement
Nursing/Other	0.83	0.82	18%	~82% less data
Radiology	0.73	0.71	15%	~85% less data
Nursing	0.70	0.78	36%	+0.08 AUC
ECG	0.61	0.58	15%	~85% less data
Physician	0.77	0.78	38%	+0.01 AUC
Echo	0.73	0.74	27%	+0.01 AUC

- Active matches or exceeds passive in most categories
- Nursing** benefits most from selective sampling
- Strong performance using only **15–38%** of the data

References

- [1] Johnson et al. MIMIC-III, 2016
- [2] Yang et al. GatorTron, 2022
- [3] Settles, Active Learning Survey, 2009
- [4] Ertekin et al., CIKM 2007
- [5] Lin et al., MICCAI 2024
- [6] Alsentzer et al. Clinical BERT, 2019
- [7] He & Garcia, Learning from imbalanced data, 2009
- [8] Knaus et al. APACHE II, 1985
- [9] Lyu et al., Multimodal Transformer, 2023
- [10] Moles et al., Data Augmentation and Active Learning in Imbalanced Datasets, 2024