Adapting TimelyGPT Model for Patient Laboratory Test Value Forecasting

Jiacheng Zhou[1], Yanxuan Yu[1], Julien W. Lee[2], Andrew Laine[1,3], Michael Sang Hughes[4,5,*],
[1]Department of Biomedical Engineering, Columbia University, New York, USA
[2]Department of Applied Mathematics, Columbia University, New York, USA
[3]Department of Radiology, Columbia University, New York, USA
[4]Department of Medicine, Columbia University Irving Medical Center, New York, USA
[5]Department of Hematology & Oncology, Columbia University Irving Medical Center, New York, USA
Email: jz3864@columbia.edu, yy3523@columbia.edu, wl2836@columbia.edu, al418@columbia.edu, mh4266@cumc.columbia.edu

***Abstract*—Forecasting the trajectory of patient laboratory values remains a challenge due to irregular sampling of measurements and differences in reference ranges between individuals. This ongoing study applies TimelyGPT, a transformer-based generative forecasting model, to predict patients' future laboratory test values based on their past records using the MIMIC-IV dataset. When predicting WBC counts, the model generated prediction errors distributed around zero, demonstrating a low bias and short-term forecasting ability. Incorporating medication administration as an additional input further shifted the error distribution toward zero and produced a more compact spread, indicating improved accuracy. Preliminary data also indicate that the model can predict sparsely sampled NT-proBNP values over time with small deviations, suggesting its potential to predict long-term laboratory test values of patients. These early results highlight the feasibility of adapting TimelyGPT model for patient-specific laboratory test values prediction and motivate work with larger cohorts, targeted data augmentation methods, and richer medication features to enhance model stability, accuracy, and generalizability.**

Keywords-electronic health record; longitudinal clinical time-series; laboratory value forecasting; generative transformer model.

I. Introduction

Patient Electronic Health Records (EHRs) contain extensive information that can be used for predictive modeling. Medical Information Mart for Intensive Care (MIMIC)-IV is a large-scale public dataset with massive patient EHR information [1]. Since the introduction of machine learning and deep learning, extensive research has been conducted to predict patient clinical conditions [2]–[4]. Recently, a model named Timely Generative Pre-trained Transformer (TimelyGPT) was introduced. This model can capture both trending and periodic features of time-series data through an extrapolatable position embedding, enabling long-term patient healthcare data forecasting [5]. The TimelyGPT model was trained on two large-scale patient EHR datasets: the Sleep European Data Format (EDF) database and the Population Health Record (PopHR) database. It is capable of predicting continuous biological signals over a short time period and can also forecast new diagnosis codes for patients based on irregularly sampled medical records. However, the study did not assess the TimelyGPT model's ability to predict the exact values of irregularly scheduled laboratory tests, which are critical for decision-making in clinical settings.

White Blood Cell (WBC) count and N-Terminus pro-Brain Natriuretic Peptide (NT-proBNP) are two fundamental markers for health in numerous diseases. In clinical practice, WBC count and NT-proBNP change from a measured baseline characterizes severity of clinically apparent inflammation and heart failure, respectively. Both assays are used to guide immediate therapy [6]. Prediction of NT-proBNP could substantially improve short- and long-term prognostication in patients with cardiovascular disease.

This study proposes to extend the current TimelyGPT framework for predicting common and sparse irregularly sampled patient laboratory values, using WBC count and NT-proBNP as proof-of-principle parameters.

While this study demonstrates the feasibility of forecasting irregularly sampled laboratory values using the adapted TimelyGPT model, several limitations should be acknowledged. First, the evaluation of WBC count prediction is conducted on a subset of patients using only data from their first six days. Therefore, the generalizability of the findings to longer time horizons remains unknown. Second, predictions of NT-proBNP values are learned and generated at discrete timestamps rather than over a continuous timeline. Finally, although the model exhibits stable predictive performance, current error margins are not yet sufficient for direct clinical deployment.

The remainder of this paper is organized as follows. Section II describes the dataset, preprocessing steps, general structure of the TimelyGPT model, and adaptations made for irregularly sampled laboratory value prediction. Section III presents the experimental results, including model performance in forecasting WBC counts with and without medication inputs, as well as NT-proBNP values. Section IV discusses the clinical implications of these results, existing methods for modeling patient EHR data, and planned future work. Finally, Section V concludes the main findings and outlines directions for future research.

II. Methods

Dataset and Preprocessing. The hospital data from the deidentified MIMIC-IV v3.1 dataset were used for model training [7]. For NT-proBNP prediction, patients were filtered to ensure an adequate length of stay (more than six timestamps) and a sufficient number of laboratory tests (at least four NT-proBNP measurements) for inclusion. For WBC count

prediction, patients were filtered to ensure an adequate length of stay (more than six relative days) and sufficient laboratory test coverage (at least four WBC count measurements).

In addition to laboratory measurements, medication administration records were extracted from the electronic medication administration record (eMAR) in MIMIC-IV v3.1 dataset. A set of 98 WBC count-influencing medications was identified based on their potential association with changes in WBC counts. For each input day, a binary medication indicator was constructed to denote whether a patient received any of the selected medications on that day. The train/validation/test splits are performed at the patient level using a 80/10/10 ratio, with no patient overlap.

Model. The TimelyGPT irregularly sampled time series algorithm was extended, and its data pipeline was modified to accommodate raw patient laboratory records (Figure 1). As illustrated in Figure 1, patient laboratory measurements are first normalized using patient- and label-specific reference ranges to ensure the stability of model training and the meaningfulness of results:

$$\text{normalized value} = \frac{\text{true value} - \text{reference range mean}}{\text{reference range width}}$$

where ***reference range width*** $=$ ***upper reference limit*** $-$ ***lower reference limit***. Each laboratory test is then transformed into a unified embedding composed of: a token embedding representing the lab test identity, a value projection encoding the normalized measurement value, and a timestamp projection capturing the time at which the test was obtained. This design allows the model to encode both laboratory measurements and their sampling timestamps, thereby capturing associated temporal dynamics.

Following the embedding construction, a start-of-sequence token is prepended, and the generated sequence is processed by stacked generative decoder layers. Each decoder layer combines multi-scale retention with temporal convolution modules, allowing the model to capture both long-range dependencies across patient trajectories as well as local temporal patterns. Finally, a feed-forward laboratory value forecasting head projects the learned representations to predict the laboratory test value at the target time point. The model uses various laboratory test values as historical information and predicts possible values for the given label at each predicted time point as output.

III. Results

TimelyGPT's predictive ability was first evaluated on WBC counts, a frequently measured laboratory parameter. The model was trained to forecast patient WBC counts in the next three days using laboratory values from the past three days. Limited to the initial six days of each patient's records, the dataset yielded 6,396 valid sequences from 13,832 patients who met the inclusion criteria. Sequences were constructed only from patients with sufficient daily WBC count measurements within the selected time window. The histogram shows that the normalized prediction errors clustered around zero, with a mean of 0.62 and a standard deviation of 0.97. Most errors fall within the range of -1 to 2, exhibiting a nearly symmetric distribution, which indicates the model's ability to predict short-term WBC count values. After incorporating WBC count-influencing medications administration as a time-resolved binary input, the error distribution further shifted towards zero and became more compact, with a reduced mean error of 0.23. These changes in the error distribution reflect improved accuracy and reduced bias, indicating that medication administrations provide meaningful information that helps the model capture WBC trajectories more effectively (Figure 2). Overall, the integration of medications contributes to more stable and accurate predictions of laboratory values.

Following validation, the model was then trained and evaluated on NT-proBNP values. Because NT-proBNP is measured infrequently in the MIMIC-IV dataset, there are insufficient sequences to train a model for predicting values on consecutive days. The first 1800 relative days of each patient's records where available were used during training. Predictions for the future three timestamps were generated based on the information from the preceding three timestamps, and adequate model convergence was observed (Figure 3). The distribution of errors normalized by reference range is centered around zero, demonstrating minimal systematic bias in the model predictions (Figure 4). Among the nearly 500 predictions, most errors fall between the -0.5 to 0.5 reference range units with few extreme errors, indicating stable and reliable model performance. To complement the histogram-based analysis, we report a standard quantitative metric for NT-proBNP prediction, which achieves a root mean squared error (RMSE) of 0.1934 across all test sequences when computed on reference-range-normalized values. Two example sequences illustrate how the model forecasts future values based on historical data (Figure 5).

IV. Discussion

These preliminary results demonstrate that the adapted TimelyGPT, when trained on healthcare data from MIMIC-IV, is capable of predicting sequential patient-specific WBC count values in the future**.** TimelyGPT can also accommodate the relatively sparse clinical data with irregular time intervals found in NT-proBNP values to make similarly accurate sequential predictions. This study proposes that TimelyGPT can forecast trajectories of common and sparse laboratory parameter values that affect patient care.

In standard clinical practice and across the medical field, a physician synthesizes available data from clinical evaluation and laboratory testing to generate an individualized prognosis, or hypothetical trajectory of disease, for a patient. Based on this prognosis, the physician then recommends a course of action. Accurate prediction of future laboratory values is thus critical to basic medical decision-making, and has a substantial impact on patients' lives [8]. Machine learning and deep learning models have in recent years augmented prognostication in multiple conditions with available data [9]–[11]. However, short- and long-term prognostication in

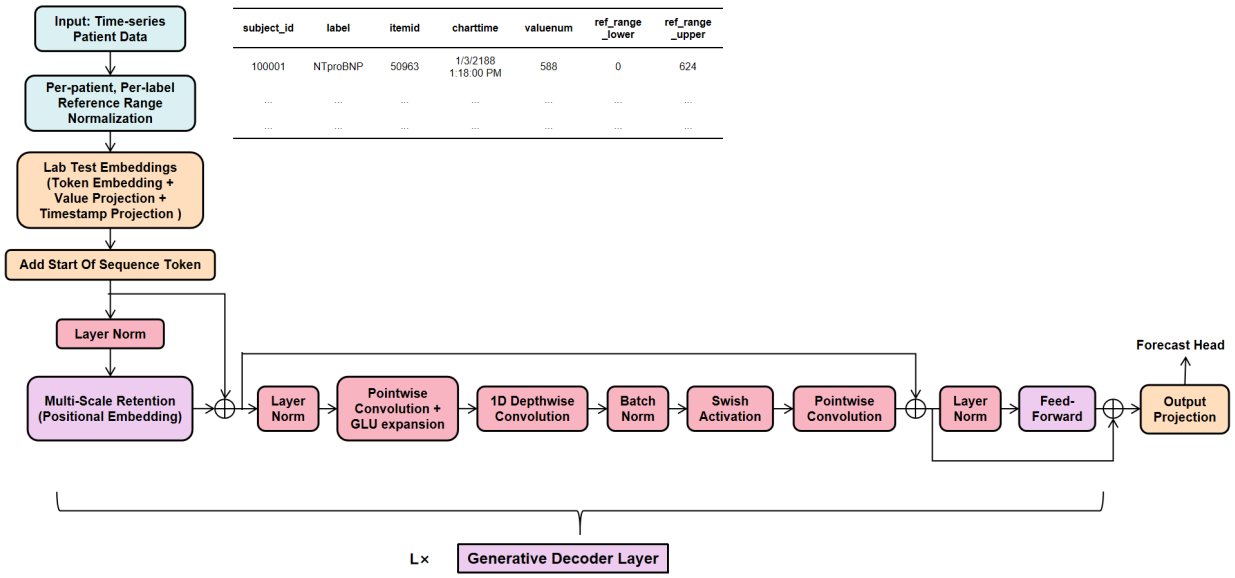

Figure 1. Overview of the adapted TimelyGPT architecture.

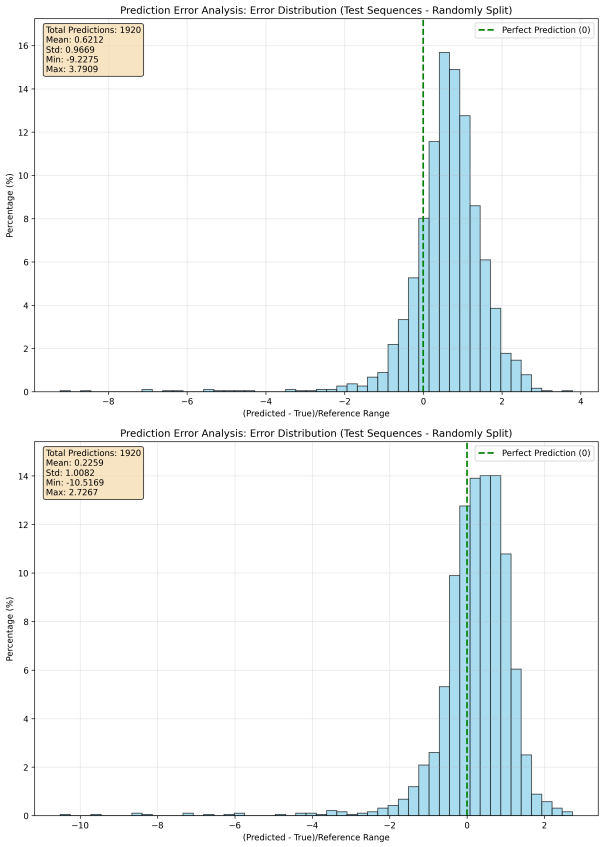

Figure 2. Error distribution of all test sequences for WBC with (bottom) and without (top) medication input.

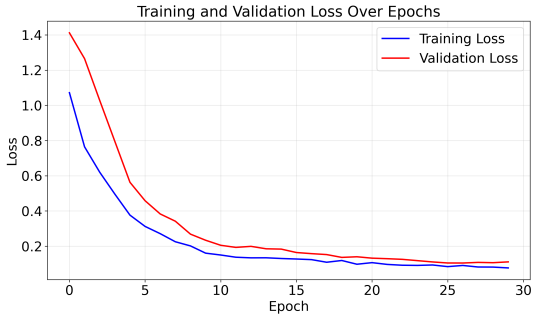

Figure 3. Loss trend over epochs when predicting NT-proBNP values using the first 1800 relative days of each patient as input.

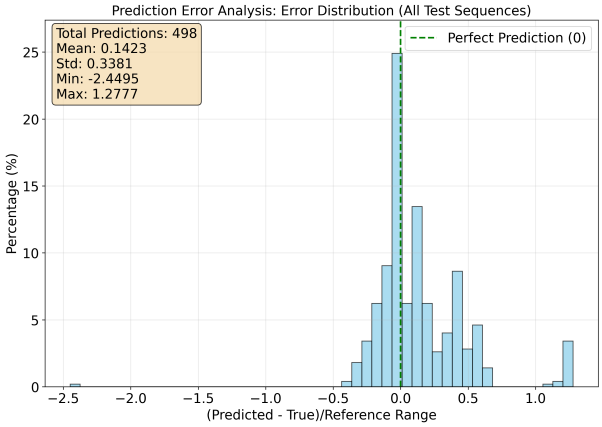

Figure 4. Error distribution of all test sequences for NT-proBNP.

numerous clinical scenarios remains a major challenge, often due to data sparsity and heterogeneous time intervals.

WBC count and NT-proBNP are two archetypal physiologic parameters which vary with disease conditions and severity. In the short term, prediction of WBC count and NT-proBNP over

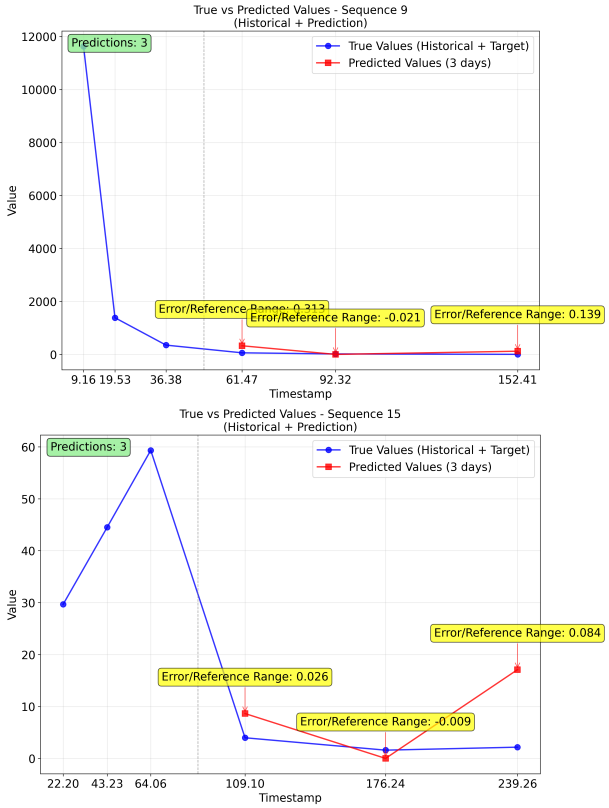

Figure 5. Plots of forecast values versus true values of NT-proBNP.

days in the inpatient setting could allow physicians to estimate date of discharge more accurately, streamlining patient flow and hospital operations. In the long term, numerous serious pathologic conditions both common and rare, from coronary artery disease to genetic cardiomyopathies, can be accompanied by progressive worsening of NT-proBNP in particular over time. The ability to predict accurate NT-proBNP values over extended time would improve individualized prognostication and allow for refinement of currently employed risk stratification systems [12].

Many efforts have been made to handle EHR data with irregular time intervals and data missingness. Traditional methods, including recurrent neural networks such as Gated Recurrent Unit (GRU)-based models and time-aware Long-Short Term Memory (LSTM) networks, incorporate decay functions in hidden states to model temporal dependencies [13][14]. Graph-guided neural networks, such as RAINDROP, construct a separate sensor graph for each sample and utilize graph neural networks to capture time-varying dependencies across variables [15]. However, these models rely on locally structured temporal information and are typically evaluated on patient Intensive Care Unit (ICU) data, where the median length of stay is about only two days, which limits their ability to perform long-term prediction tasks. More recently, transformer-based architectures have been explored to handle long sequences from healthcare data. The Perceiver architecture has demonstrated the ability to capture global temporal dynamics through cross-attention for continuous time modeling when combined with a neural Ordinary Differential Equation (ODE) module [16]. TimelyGPT further advanced this by incorporating retention mechanisms directly within its architecture, enabling the modeling of continuous temporal dependencies in the long term without requiring additional differentiation computations. By extending TimelyGPT, the ultimate goal is to forecast patient laboratory values in both short- and long-term for inpatient and outpatient settings.

While TimelyGPT can predict even sparse laboratory parameters such as NT-proBNP with adequate accuracy, such margins of error at this time are still too wide for clinical use. In addition, current results should be interpreted as proof-of-concept for predicting irregularly sampled laboratory values. In this study, NT-proBNP values are predicted at future observed timestamps rather than fixed temporal intervals, and the elapsed time can be highly variable across patients.

Thus, we plan to implement three strategies to further improve the model's performance. Firstly, further model training will be conducted on outpatient data from the Integrating Numerous Sources for Prognostic Evaluation of Clinical Timelines (INSPECT) database, which has 26,795 additional instances of NT-proBNP measurements [17]. Incidentally, considerations of outpatient versus inpatient laboratory parameter prediction are intended, which will further help our model's generalizability. If data are still insufficient, novel data augmentation methods will be further explored to acquire more NT-proBNP values, using existing algorithms such as Generative Adversarial Networks for Mixed-type EHR data (EHR-M-GAN) or building on our recent work with Synthetic Minority Over-sampling Technique with Adversarial Filtering (AF-SMOTE) [18][19]. Simultaneously, the administration of specific medications that can affect NT-proBNP value, such as diuretics, will be integrated as an additional input. A significant reduction in model output error after incorporation is expected.

V. Conclusion and Future Work

In this paper, we extended the TimelyGPT framework to forecast laboratory test values from patient EHRs. Using data from MIMIC-IV, we evaluated the model on both frequently measured WBC counts and sparsely sampled NT-proBNP. Overall, these findings demonstrate that adapting the TimelyGPT model for single-label and irregularly-sampled healthcare parameter forecasting is feasible. Incorporating medication administrations might further enhance the model's ability to capture laboratory trajectories, as these inputs provide temporal cues that reflect therapeutic effects. Additional work will be done to expand the training datasets to include a broader population, introduce data augmentation methods to improve the representations of rare laboratory values, and incorporate medication effects to enhance model performance. Collectively, these strategies are expected to further improve the accuracy, robustness, and generalizability of TimelyGPT-based predictions.

REFERENCES

[1] A. Johnson *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023. DOI: 10.1038/s41597-022-01899-x.

[2] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, p. 18, 2018. DOI: 10.1038/s41746-018-0029-1.

[3] J. Kundrick *et al.*, "Machine learning applied to wearable fitness tracker data and the risk of hospitalizations and cardiovascular events," *American Journal of Preventive Cardiology*, vol. 22, p. 101 006, 2025. DOI: 10.1016/j.ajpc.2025.101006.

[4] A. Amirahmadi, M. Ohlsson, and K. Etminani, "Deep learning prediction models based on ehr trajectories: A systematic review," *Journal of Biomedical Informatics*, vol. 144, p. 104 430, 2023. DOI: 10.1016/j.jbi.2023.104430.

[5] Z. Song *et al.*, "Timelygpt: Extrapolatable transformer pre-training for long-term time-series forecasting in healthcare," *Health Information Science and Systems*, vol. 13, no. 1, p. 64, 2025. DOI: 10.1007/s13755-025-00384-0.

[6] P. Jourdain *et al.*, "Plasma brain natriuretic peptide-guided therapy to improve outcome in heart failure:the STARS-BNP multicenter study," *Journal of the American College of Cardiology*, vol. 49, no. 16, pp. 1733–1739, 2007. DOI: 10.1016/j.jacc.2006.10.081.

[7] A. Johnson *et al.*, "Mimic-iv," *PhysioNet*, Oct. 2024, Version 3.1. DOI: 10.13026/kpb9-mt58. [Online]. Available: https://doi.org/10.13026/kpb9-mt58.

[8] J. M. Thomas, L. M. J. Cooney, and T. R. Fried, "Prognosis as health trajectory: Educating patients and informing the plan of care," *Journal of General Internal Medicine*, vol. 36, no. 7, pp. 2125–2126, 2021. DOI: 10.1007/s11606-020-06505-7. [Online]. Available: https://doi.org/10.1007/s11606-020-06505-7.

[9] D. Ramamoorthy *et al.*, "Identifying patterns in amyotrophic lateral sclerosis progression from sparse longitudinal data," *Nature Computational Science*, vol. 2, no. 9, pp. 605–616, 2022. DOI: 10.1038/s43588-022-00299-w. [Online]. Available: https://doi.org/10.1038/s43588-022-00299-w.

[10] G. A. Kwong *et al.*, "Synthetic biomarkers: A twenty-first century path to early cancer detection," *Nature Reviews Cancer*, vol. 21, no. 10, pp. 655–668, 2021. DOI: 10.1038/s41568-021-00389-3. [Online]. Available: https://doi.org/10.1038/s41568-021-00389-3.

[11] S. N. Naik *et al.*, "Unsupervised airway tree clustering with deep learning: The multi-ethnic study of atherosclerosis (mesa) lung study," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, p. 10. DOI: 10.1109/isbi56570.2024.10635651. [Online]. Available: https://doi.org/10.1109/isbi56570.2024.10635651.

[12] A. Cai *et al.*, "Heart stress and blood pressure management in older adults: Post hoc analysis of the aspree trial," *Circulation*, Oct. 2025, Epub ahead of print. DOI: 10.1161/CIRCULATIONAHA.125.076263.

[13] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, p. 6085, 2018. DOI: 10.1038/s41598-018-24271-9.

[14] I. M. Baytas *et al.*, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 65–74, ISBN: 9781450348874. DOI: 10.1145/3097983.3097997. [Online]. Available: https://doi.org/10.1145/3097983.3097997.

[15] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, *Graph-guided network for irregularly sampled multivariate time series*, 2022. arXiv: 2110.05357 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2110.05357.

[16] V. K. Chauhan *et al.*, "Continuous patient state attention model for addressing irregularity in electronic health records," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 117, 2024. DOI: 10.1186/s12911-024-02514-2.

[17] S. Huang *et al.*, *Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis*, 2023. arXiv: 2311.10798 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2311.10798.

[18] J. Li, B. J. Cairns, J. Li, and T. Zhu, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications," *npj Digital Medicine*, vol. 6, p. 98, 2023. DOI: 10.1038/s41746-023-00834-7.

[19] Y. Yu, M. S. Hughes, J. Lee, J. Zhou, and A. F. Laine, *Boundary-aware adversarial filtering for reliable diagnosis under extreme class imbalance*, 5 pages, 3 figures. Submitted to IEEE ISBI (under review), 2025. DOI: 10.48550/arXiv.2511.17629. arXiv: 2511.17629 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2511.17629.