ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN NGHIÊN CỨU 3

Dự đoán bệnh tan máu bằng mô hình học máy và học sâu

Phạm Anh Minh
minh.pa194802@sis.hust.edu.vn

Ngành: Công nghệ thông tin

Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang ______________

Chữ kí GVHD

Khoa: Kỹ thuật máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 06/2024

TÓM TẮT NỘI DUNG ĐỒ ÁN

Sepsis is a severe complication arising from an infection. If not treated promptly, it can result in organ failure and death. Therefore, early detection and treatment of sepsis can potentially save many lives. However, their effectiveness often depends on the awareness and acceptance of these procedures. In this study, i implement a sepsis check based on the widely accepted Sepsis-3 guidelines. My implementation achieved an F-score of up to 0.874. Alongside the rule-based approach for early sepsis detection, i also employ the existing data-driven transformer-based STraTS model (Tipirneni and Reddy, 2021) for time-series forecasting to support sepsis checks and directly predict sepsis using 24-hour patient data in a fully data-driven setup. Furthermore, i aim to enhance the mono-modal STraTS model by incorporating a clinical text embedding module to enable multi-modal learning. Both the original STraTS model and my refined STraTS+Text model performed well in forecasting (with a masked MSE of approximately 5.24) and classification tasks (with a ROC-AUC of approximately 0.89).

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

CHAPTER 1. INTRODUCTION

Sepsis occurs when the body's immune response to an infection becomes dysregulated, leading to systemic inflammation. It is a leading cause of death in Intensive Care Units (ICU). Early detection is crucial for patient survival (Rudd et al., 2020). To identify septic patients from their clinical data, Singer et al. (2016) and Reyna et al. (2019) present slightly different rule-based guidelines focusing on suspected infections and clinical criteria for life-threatening organ dysfunction. Singer et al.'s (2016) guidelines were developed as an in-hospital tool to assess patient condition.These guidelines can be applied to observed data and to forecast time-series values, potentially allowing earlier identification and prevention of sepsis. I implement a rule-based sepsis check based on the widely accepted Sepsis-3 guidelines to enable early sepsis prediction.

Additionally, I refine an existing Self-supervised Transformer for Time-Series (STraTS) model (Tipirneni and Reddy, 2021) for time-series forecasting and 24-hour sepsis prediction. The STraTS regression model forecasts time-series values following each observation window to support the sepsis check, while the STraTS classification model predicts sepsis using 24-hour patient data in a fully data-driven setup. I enhance the original STraTS architecture, which only takes continuous physiological features as input, by integrating a clinical text embedding module based on Clinical BERT (Alsentzer et al., 2019) to encode 1.4 million clinical notes from patients in my MIMIC-IV data.Both models (STraTS and STraTS+Text) perform well in forecasting and classification tasks, achieving a masked Mean Squared Error (MSE) of approximately 5.24 and a ROC-AUC of approximately 0.89.

My rule-based sepsis check achieved an F-score of up to 0.874 without using features predicted by the STraTS forecasting model. While introducing predicted values from the STraTS forecasting model did not improve the rule-based sepsis check's performance, these predictions helped address data sparsity, enabling the rule-based check to identify septic patients whose clinical data alone would not have been sufficient for accurate classification.

CHAPTER 2. RELATED WORK

The study by Torio and Andrews (2013) highlighted that in 2011, sepsis was the most costly condition treated in U.S. hospitals, accounting for 5.2% of total hospitalization costs. In a global context, Rudd et al. (2020) found that in 2017, sepsis caused 19.7% of all global deaths, with a higher incidence in low- and middle-income countries. Despite existing guidelines for identifying and treating septic patients, such as those proposed by Singer et al. (2016), Kissoon (2014) argued that factors like critical staff shortages, failure to identify sepsis promptly, and limited availability of laboratory tests complicate their implementation on-site.Historically, the diagnosis and understanding of sepsis heavily relied on the Systemic Inflammatory Response Syndrome (SIRS) criteria (Bone et al., 1992). However, the reliability of SIRS has been questioned, leading to the development of the Sepsis-related Organ Failure Assessment (SOFA) score (Vincent et al., 1996) as a more robust tool.

With the increasing use of electronic health records (EHR) systems, clinical time-series data, and digital clinical notes, machine learning models have been employed for prediction tasks in medicine, including sepsis and mortality prediction (Wang et al., 2022; Tipirneni and Reddy, 2021). Earlier approaches used linear dynamical systems (LDS) and Gaussian processes (GP) to model clinical time-series data (Liu et al., 2013; Liu and Hauskrecht, 2015). More recent advancements include models like BioBERT and ClinicalBERT, pretrained on medical text data, to analyze medical articles and clinical notes (Lee et al., 2020; Alsentzer et al., 2019).

Recent studies have moved beyond basic data modeling to tackle specific predictive tasks. Wang et al. (2022) proposed a multi-modal learning approach for early sepsis prediction, integrating clinical notes and continuous physiological features into a transformer-based binary classification model. This model demonstrated strong performance on datasets like MIMIC-IV and eICU-CRD for predicting whether patients develop sepsis shortly after ICU admission. In contrast, Tipirneni and Reddy (2021) introduced a two-step transformer-based method for mortality prediction, incorporating a regression model to forecast time-series values and a classification model to predict sepsis labels based on patient data.

CHAPTER 3. SEPSIS-3 IMPLEMENTATION

3.1 Suspected Infection

According to the third International Consensus Definitions for Sepsis and Septic Shock (Singer et al., 2016), sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection. A key indicator for septic patients is a suspected infection. With available data, a suspected infection is identified by orders for blood cultures and antibiotics. The Singer et al. (2016) guideline specifies a particular time frame in which antibiotics and cultures must be ordered: if antibiotics are administered first, blood cultures need to be taken within 24 hours, and if blood cultures are taken first, antibiotics must be administered within 72 hours. This time frame is referred to as the suspicion window.

In contrast, Reyna et al. (2019) require antibiotics to be administered for at least 72 consecutive hours to be considered a suspected infection. In this case, the first administration of antibiotics is compared to blood culture orders similarly to Singer et al. (2016). This distinction significantly impacts implementation, as the sparsity of available data and the ability to identify consecutive administrations of antibiotics present a non-trivial challenge.

3.2 Criterion for life- threatening organ dysfunction

Another key indicator for septic patients is life-threatening organ dysfunction. Several criteria can be employed to identify this, but both Singer et al. (2016) and Reyna et al. (2019) recommend using the Sequential [Sepsis-related] Organ Function Assessment (SOFA) score (Vincent et al., 1996), which considers a variety of clinical and laboratory variables. According to Singer et al. (2016), a patient's SOFA score should be computed hourly. The time of SOFA is when a patient's score reaches two or higher, assuming an initial value of zero if no prior organ dysfunction is known. Reyna et al. (2019) suggest that the time of SOFA is when there is an increase of two points compared to the previous 24 hours. If this time of SOFA is within 48 hours before or 24 hours after a suspected infection, the patient is considered to have developed sepsis according to Singer et al. (2016). Reyna et al. (2019) are more stringent, allowing the time of SOFA to be within only 24 hours before and 12 hours after the suspected infection. Additionally, they treat the earlier of the two times as the onset of sepsis. This time period between suspected infection and time of SOFA is referred to as the sepsis window in the implementation.

3.3 Implementation

To compute either the Singer et al. (2016) or Reyna et al. (2019) checks, the clinical features described in Tables 11 and 10 should be included in the patient data, with particular emphasis on antibiotics and blood culture features. Given the nature of these rule-based guidelines, the absence of either the antibiotics or blood culture feature will always result in a negative sepsis label, as these features are essential for suspecting an infection.

3.4 About preprocessing and running the sepsis check

The sepsis check includes several utility functions to process the output from Tipirneni and Reddy (2021). Before starting an experiment, decisions need to be made regarding whether the antibiotics feature should be imputed by forward filling, which strategy should be employed, and what the sepsis and suspicion windows should be. The necessary features are then extracted from the preprocessed patient data.The preprocessing by Tipirneni and Reddy (2021) includes normalization using the formula:

$$\text{normalized} = \frac{\text{value} - \text{mean}}{\text{std}} \qquad (1)$$

Consequently, numerical features are renormalized and then aggregated per hour. Next, the data is imputed by forward filling. The features for blood cultures, text, mechanical ventilation, and catecholamines are excluded by default, and the antibiotics feature is filled based on the initial decision. Finally, the features are cast to their correct types.

3.5 SOFA

Before the SOFA scores can be computed, the Glasgow Coma Scale (GCS) is derived from its components. Additionally, the mean arterial pressure (MAP) is estimated using diastolic and systolic blood pressures according to Demers and Wachs (2019) with the following formula:

$$\text{MAP} = \text{DBP} + \frac{\text{SBP} - \text{DBP}}{3} \qquad (2)$$

Once these calculations are made, the SOFA score is computed for each hour. Following this, the time of SOFA can be identified according to the guidelines.

3.6 Suspected Infection and Sepsis Classification

The features for antibiotics and blood cultures are checked according to the set strategy and suspicion window. If the conditions are met, the earlier time is considered the time of suspicion, which is then compared to the time of SOFA under the constraints of the sepsis window. This part of the sepsis check is critical. If either blood cultures or antibiotics are not reported in the patient data, the patient is considered not to have developed sepsis according to the guidelines.During the evaluation of initial experiments, it became evident that the primary reason for erroneous patient classifications was the lack of a time of suspicion. Initially, it was suspected that the suspicion window was the problem. Increasing the suspicion window to up to ten days increased the number of suspected infections, but they were suspected too late and missed the sepsis window. This issue is more serious for Reyna et al. (2019), as antibiotics need to be administered for 72 consecutive hours. If a hospital stay is less than 72 hours, there can be no sepsis. Furthermore, if the antibiotics data is too sparse, there can be no sepsis without forward-filling the feature. Even with forward-filling, there can be no sepsis if the patient did not stay at least 72 hours after the first administration of antibiotics.

To address this problem, two additional strategies were implemented. While the Sepsis-3 strategy uses the first time of blood cultures and the first time of antibiotics within the supplied suspicion window to compute the time of suspicion, the 'catchsus' strategy considers all times blood cultures were taken and all times antibiotics were administered. It checks if any of the possible combinations of the time of antibiotics and time of blood cultures fall within the specified suspicion window. For each combination that falls within the suspicion window, the earlier time is considered the time of suspicion. This can yield multiple times of suspicion, which are then compared to the time of SOFA and the sepsis window to generate a sepsis label. The 'grouped' strategy takes this approach further by also considering multiple times of SOFA. As a result, multiple times of suspicion and multiple times of SOFA are considered when generating a sepsis label. Unfortunately, even though the 'catchsus' and 'grouped' strategies outperformed the standard strategies on unfiltered data—which contains many patients that are impossible to correctly predict for the rule-based guidelines due to missing antibiotics or blood culture features—the standard strategies performed better on patients where a positive sepsis label was possible. This indicated that the 'catchsus' and 'grouped' strategies were benefiting from the data distribution rather than being better strategies.

3.7 Utilizing Time-Series Forecasting

Next to the potential advantage of predicting the onset of sepsis before observing the features that indicate sepsis, another benefit of using time-series forecasting is the ability to forecast important features. In this case, antibiotics and blood cultures are binary features, while Tipirneni and Reddy (2021) outputs continuous values.

To interpret these continuous values, a threshold is set to assign a binary label (e.g., positive/negative) based on whether the value is above or below the threshold. The method to determine this threshold can be improved, currently involving a combination of clustering and iterative testing.The experiments conducted for this research paper combine observed data with one-hour-ahead forecasting output based on that observed data. For each observation window ranging from 20 to 120 hours in steps of four, the resulting concatenated data is used as input for the sepsis check. This approach allows for evaluating how well forecasting aids in predicting sepsis onset using different time frames of data aggregation.

CHAPTER 4. DATA

4.1 MIMIC-IV

The MIMIC-IV dataset is a comprehensive collection of electronic health records (EHRs) from patients admitted to intensive care units (ICUs) in the United States. It contains de-identified data on over 400,000 patients, including demographics, vital signs, laboratory results, medications, and clinical outcomes. The dataset is designed to support research in critical care medicine, with a focus on improving patient outcomes and reducing healthcare costs. MIMIC-IV is widely used by researchers and clinicians to develop and evaluate machine learning models for predicting patient mortality, sepsis, and other ICU-related complications. For a detailed list of tables, you can refer to Table 4 or explore the data exploration section of related code.

Model	ROC-AUC	PR-AUC	min(Re,Pr)
STraTS	0.891 ± 0.003	0.500 ± 0.009	0.507 ± 0.100
STraTS + Text	0.889 ± 0.002	0.491 ± 0.008	0.492 ± 0.008

Bảng 4.1: Sepsis prediction performance on MIMIC-IV dataset. The results show mean and standard deviation of the metrics after repeating the experiment 10 times by sampling 50% labeled data each time.

4.2 Sepsis Label Annotation

My project is built upon the STraTS framework (Tipirneni and Reddy, 2021), which was originally designed for mortality prediction requiring a mortality label. In my application focusing on sepsis prediction, i adapted a sepsis detection method discussed earlier. To identify patients with sepsis, i used ICD9 codes extracted from the diagnosis table, specifically selecting 23 codes related to sepsis as detailed in Table 7 of Appendix B. After preparing the data, each hospital admission ID was assigned a sepsis label. Additionally, i filtered out patients admitted with sepsis from my dataset by matching strings in the admissions table of MIMIC-IV. This step was crucial as these patients would not contribute relevant information for my forecasting model.

4.3 My Data

In my dataset sourced from MIMIC-IV, i included 5,288 patients identified with sepsis, comprising 9.2% of the total, alongside 51,994 patients without sepsis. The data was partitioned into training, validation, and test sets at a 64:16:20 ratio based on patient counts (see Table 1).

Data	N	n	Non-septic patients	Septic patients
Train	26452	33191	2124	3360
Valid	6594	8358	551	904
Test	8296	10445	635	1024

Bảng 4.2: Number of septic/non-septic patients/ICU stays in train/validation/test data.

Additionally, my dataset incorporates 1,407,430 clinical notes from MIMIC-IV, enriching the dataset beyond 133 physiological features listed in Appendix A. Table 2 provides detailed statistics on these clinical notes associated with the patients in my study.

	Avg.	Max.	Min.
String length	1673	55728	3
Num. tokens	316	11336	0

Bảng 4.3: String length and token counts in clinical notes included in my data.

4.4 Clinical Notes Preprocessing

I preprocess clinical notes using standard methods for clinical text cleaning. This involves removing stop words and special characters, as well as normalizing text case. To prevent any potential label leakage in the STraTS classification task, i also filter out sentences that contain the terms ’sepsis’ or ’septic’.

5.1 STraTS

I employ the STraTS model (Tipirneni and Reddy, 2021) for two main tasks: forecasting physiological feature values in time-series for my rule-based sepsis check, and directly predicting sepsis labels. The STraTS model uses Continuous Value Embedding to represent continuous data without traditional discretization methods like aggregation or imputation. Each data observation is encoded as a triplet comprising observed time, variable name, and variable value. These embeddings undergo transformation through a multi-head attention mechanism and a fusion self-attention module, enabling contextual learning of time-series data.

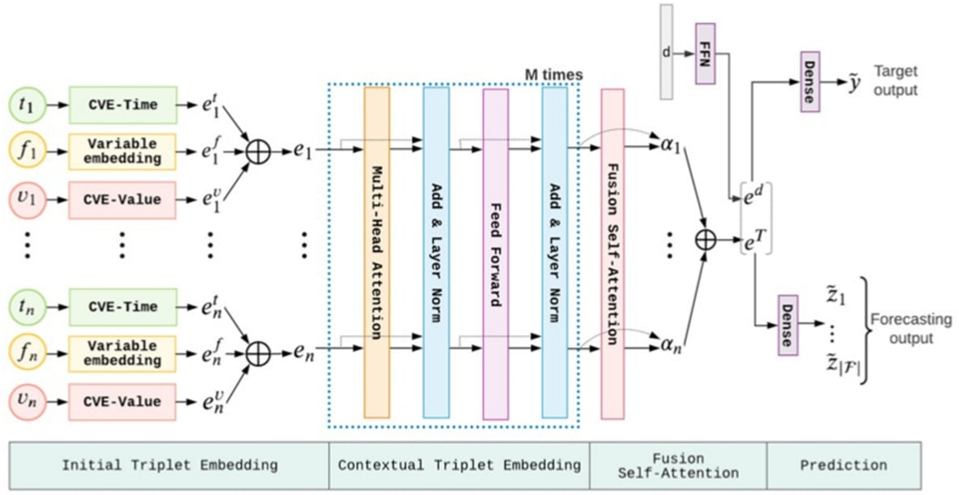
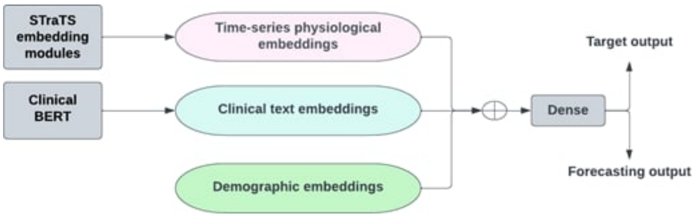


Fig. 3. The overall architecture of the proposed STraTS model. The Input Triplet Embedding module embeds each observation triplet, the Contextual Triplet Embedding module encodes contextual information for the triplets, the Fusion Self-Attention module computes time-series embedding, which is concatenated with demographics embedding and passed through a dense layer to generate predictions for target and self-supervision (forecasting) tasks.

The STraTS architecture supports both regression for time-series forecasting and binary classification for predicting state labels such as mortality or sepsis. Originally designed to handle limited labeled data in medical contexts, the regression model serves as an auxiliary task to enhance performance of the classification model during forecasting. In my study, i utilize both models to produce forecasting outputs that support my rule-based sepsis check, and to predict sepsis states over 24-hour periods in a fully data-driven approach.

5.2 STraTS + Clinical Text Embedding

I enhance the original STraTS model, which initially processes only physiological features for forecasting and classification, by integrating a clinical text embedding module. Using Clinical BERT (Alsentzer et al., 2019), i embed clinical notes to extract text features. These text features are aligned alongside time-series embeddings of physiological and demographic features, concatenated together, and passed through a dense layer to produce outputs for both forecasting and classification tasks. Figure 1 illustrates the architecture of the refined STraTS model with clinical text embedding.



Hình 5.1: STraTS + Clinical Text Embedding Architecture.

CHAPTER 6. RESULTS AND DISCUSSION

6.1 STraTS Forecasting

I trained STraTS regression models with and without clinical text embeddings to forecast physiological feature values in the two hours following observation windows, defined as $\{min(0, x - 24), x|20 \leq x \leq 124, x\%4 = 0\}$. Predictions from both regression models were obtained on test data to support the rule-based sepsis check. Evaluation was based on masked MSE (mean squared error), where a binary mask indicates whether a true value was observed in the data. Table 3 presents the masked MSE results on test and validation data for both STraTS and STraTS + Text regression models. The table illustrates that while both models performed similarly on test data, the original STraTS model without clinical text embeddings exhibited better MSE on validation data.

Model	Test MSE	Validation MSE
STraTS	5.2455	5.2048
STraTS + Text	5.2493	5.5922

Bảng 6.1: Masked MSE (mean squared error) on test and validation data for STraTS and STraTS + Text models.

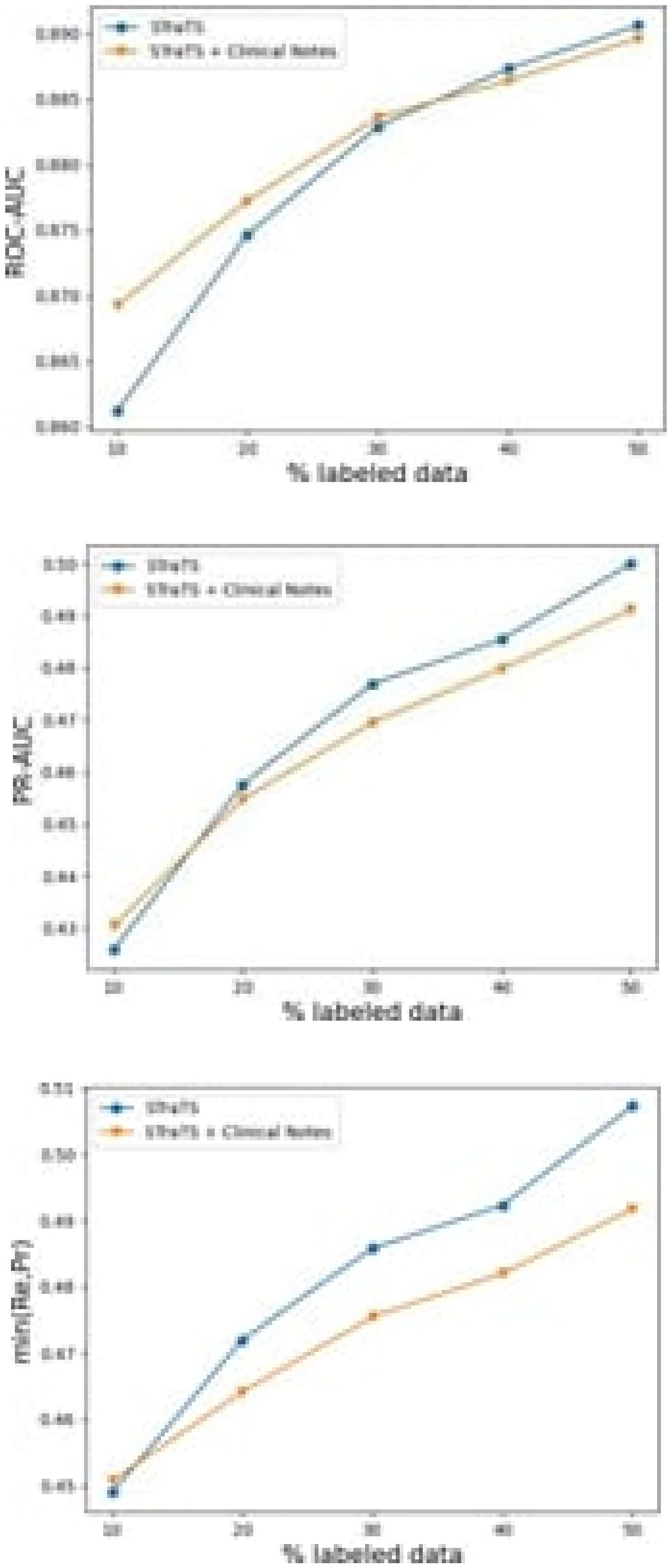
6.2 STraTS Classification

I trained STraTS classification models with and without clinical text embeddings using a random sample of 10%, 20%, 30%, 40%, and 50% labeled data from 24-hour ICU admission data. Each experiment was repeated 10 times with different randomly sampled data from the train and validation sets.

For evaluating the models' performance in the binary classification task predicting septic state, i used three metrics:

- ROC-AUC: Area under the ROC curve.
- PR-AUC: Area under the precision-recall curve.
- min(Re, Pr): The maximum of the minimum of recall and precision across all thresholds.

Figure 2 displays ROC-AUC, PR-AUC, and min(Re, Pr) scores of both models on the test dataset. The charts indicate that STraTS + Clinical Text slightly outperforms STraTS when labeled data is limited $(\leq 30\%)$ in terms of ROC-AUC score. However, STraTS generally exhibits higher PR-AUC and min(Re, Pr) scores except when only 10% of labeled data is available. STraTS + Clinical Text shows marginally better performance in scenarios with limited labeled data, while STraTS performs better when more labeled data is available.



Hình 6.1: Sepsis prediction performance on MIMIC-IV dataset for different percentages of labeled data averaged over 10 runs

6.3 Sepsis Check results

While the experiments incorporating time-series forecasting only include one additional hour of unobserved data, the results in Table 5 are somewhat less significant than anticipated. However, it is important to highlight that the additional data contributed to more accurate identification of true positives across all experiment sets, including patients initially lacking both antibiotics and blood culture features in their observed data. This suggests that STraTS could potentially mitigate issues stemming from data sparsity. Expanding the STraTS predictions to multiple hours appears to be a promising avenue for future research.

Experiment	F1 score	True
1.1 observed-only	0.874505	251
1.2 observed+forecast	0.873794	286
2.1 observed-only	0.87309	234
2.2 observed+forecast	0.873677	265

Bảng 6.2: 1: experiments were conducted with a suspicion window of 48 and 72 hours, and a sepsis window of 24 and 12 hours. 2: experiments were conducted with a suspicion window of 24 and 96 hours, and a sepsis window of 24 and 12 hours.

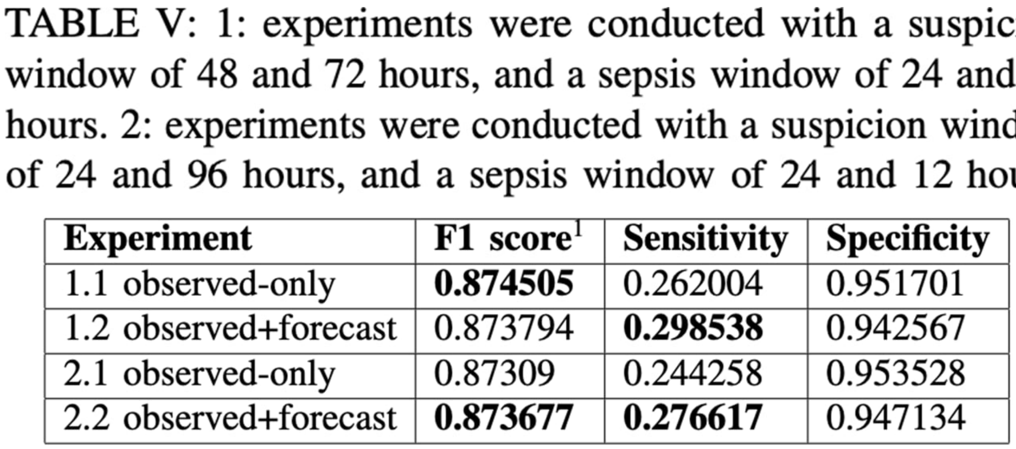


TABLE V: 1: experiments were conducted with a suspicion window of 48 and 72 hours, and a sepsis window of 24 and 12 hours. 2: experiments were conducted with a suspicion window of 24 and 96 hours, and a sepsis window of 24 and 12 hours.

Experiment	**F1 score**[1]	**Sensitivity**	**Specificity**
1.1 observed-only	**0.874505**	0.262004	0.951701
1.2 observed+forecast	0.873794	**0.298538**	0.942567
2.1 observed-only	0.87309	0.244258	0.953528
2.2 observed+forecast	**0.873677**	**0.276617**	0.947134

CHAPTER 7. CONCLUSION

I implemented a rule-based sepsis check and observed that the sparse patient data significantly impacts potential performance. While my rule-based approach demonstrated good standalone performance, integrating forecasts from the current STraTS regression model did not enhance the sepsis check in my experiments. My approach aims to address real-world complexities in sepsis treatment as outlined by Kissoon (2014).

In addition to the rule-based sepsis check, i utilized and refined the STraTS model (Tipirneni and Reddy, 2021) for time-series forecasting to support sepsis detection and direct prediction of sepsis labels using 24-hour patient data in a fully data-driven setup. I endeavored to enhance the STraTS model by incorporating a clinical text embedding module based on Clinical BERT, enabling multi-modal learning by integrating text and physiological features. Both the STraTS and STraTS+Text regression models demonstrated robust performance in forecasting time-series values, and the classification models exhibited high ROC-AUC scores for sepsis prediction using features selected in my study.

However, my current STraTS+Text model did not surpass the original STraTS model in my experiments. For future research, i plan to further refine the STraTS+Text model architecture, particularly focusing on enhancing the text embedding module to improve feature representation. Additionally, to better support the rule-based sepsis check, i aim to extend the STraTS architecture to enable forecasting over longer windows based on short and limited observational periods.

CHAPTER 8. REFERENCES

1. Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

2. Roger C. Bone, Robert A. Balk, Frank B. Cerra, R. Phillip Dellinger, Alan M. Fein, William A. Knaus, Roland M.H. Schein, and William J. Sibbald. 1992. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655.

3. Daniel J. Demers and Daliah Wachs. 2019. Physiology, mean arterial pressure.

4. Niranjan Kissoon. 2014. Sepsis guideline implementation: benefits, pitfalls and possible solutions. *Critical Care*, 18(2).

5. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

6. Zitao Liu and Milos Hauskrecht. 2015. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial intelligence in medicine*, 65(1):5–18.

7. Zitao Liu, Lei Wu, and Milos Hauskrecht. 2013. Modeling clinical time series using Gaussian process sequences. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 623–631. SIAM.

8. Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13.

9. Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth Prajwal Shashikumar, Michael Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. 2019. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48:210–217.

10. Kristina E. Rudd, Sarah Charlotte Johnson, Kareha M. Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V. Colombara, Kevin

S. Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R. Machado, Konrad K. Reinhart, Kathryn Rowan, Christopher W. Seymour, R. Scott Watson, T. Eoin West, Fatima Marinho, Simon I. Hay, Rafael Lozano, Alan D. Lopez, Derek C. Angus, Christopher J. L. Murray, and Mohsen Naghavi. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211.

11. Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810.

12. Celeste Marie Torio and Roxanne M. Andrews. 2013. National inpatient hospital costs: The most expensive conditions by payer, 2011.

13. J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. 1996. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710.

14. Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. 2022. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis.

PHỤ LỤC

A. FEATURES

Physiological Features

- ALP
- ALT
- AST
- Albumin
- Albumin 25%
- Albumin 5%
- Amiodarone
- Anion Gap
- Antibiotics
- BUN
- Base Excess
- Basophils
- Bicarbonate
- Bilirubin (Direct)
- Bilirubin (Indirect)
- Bilirubin (Total)
- Blood Culture
- CRR
- Calcium Free
- Calcium Gluconate
- Calcium Total
- Cefazolin
- Chest Tube
- Chloride
- Colloid
- Creatinine Blood

- Creatinine Urine
- D5W
- DBP
- Dextrose Other
- Dopamine
- EBL
- Emesis
- Eosinophils
- Epinephrine
- Famotidine
- Fentanyl
- FiO2
- Fiber
- Free Water
- Fresh Frozen Plasma
- Furosemide
- GCS_eye
- GCS_motor
- GCS_verbal
- GT Flush
- Gastric
- Gastric Meds
- Glucose (Blood)
- Glucose (Serum)
- Glucose (Whole Blood)
- HR
- Half Normal Saline
- Hct
- Height

- Heparin
- Hgb
- Hydralazine
- Hydromorphone
- INR
- Insulin Humalog
- Insulin NPH
- Insulin RegularInsulin largine
- Intubated
- Jackson-Pratt
- KCl
- KCl (Bolus)
- LDH
- Lactate
- Lactated Ringers
- Levofloxacin
- Lorazepam
- Lymphocytes
- Lymphocytes (Absolute)
- MBP
- MCH
- MCHC
- MCV
- Magnesium
- Magnesium Sulfate (Bolus)
- Magnesium Sulphate
- Mechanically ventilated
- Metoprolol
- Midazolam

- Milrinone
- Monocytes
- Morphine Sulfate
- Neosynephrine
- Neutrophils
- Nitroglycerine
- Nitroprusside
- Norepinephrine
- Normal Saline
- O2 Saturation
- OR/PACU Crystalloid
- PCO2
- PO intake
- PO2
- PT
- PTT
- Packed RBC
- Pantoprazole
- Phosphate
- Piggyback
- Piperacillin
- Platelet Count
- Potassium
- Pre-admission Intake
- Pre-admission Output
- Propofol
- RBC
- RDW
- RR

- Residual
- SBP
- SG Urine
- Sodium
- Solution
- Sterile Water
- Stool
- TPN
- Temperature
- Total CO2
- Ultrafiltrate
- Unknown
- Urine
- Vancomycin
- Vasopressin
- WBC
- Weight
- pH Blood
- pH Urine

Demographic Features

- Age
- Gender

B. SEPSIS CODES FROM MIMIC-IV

The positive sepsis labels in our dataset were determined using specific ICD-9 diagnosis codes extracted from the D_ICD_DIAGNOSES table. These codes include various conditions associated with sepsis, such as bacterial infections and related systemic inflammatory responses. They were employed to identify patients diagnosed with sepsis during their hospital admissions.

ICD9 Code	Short Description
0380	Streptococcal septicemia
03810	Staphylcocc septicem NOS
03811	Meth susc Staph aur sept
03812	MRSA septicemia
03819	Staphylcocc septicem NEC
0382	Pneumococcal septicemia
0383	Anaerobic septicemia
03840	Gram-neg septicemia NOS
03841	H. influenzae septicemia
03842	E coli septicemia
03843	Pseudomonas septicemia
03844	Serratia septicemia
03849	Gram-neg septicemia NEC
0388	Septicemia NEC
0389	Septicemia NOS
67020	Puerperal sepsis-unsp
67022	Puerprl sepsis-del w p/p
67024	Puerperl sepsis-postpart
67030	Puerp septc thromb-unsp
67032	Prp sptc thrmb-del w p/p
67034	Prp septc thrmb-postpart
99591	Sepsis
99592	Severe sepsis

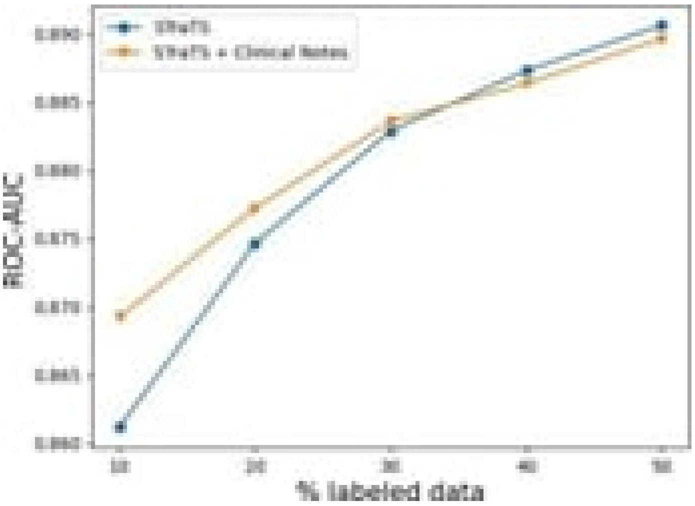
Bảng B.1: ICD9 Codes for sepsis

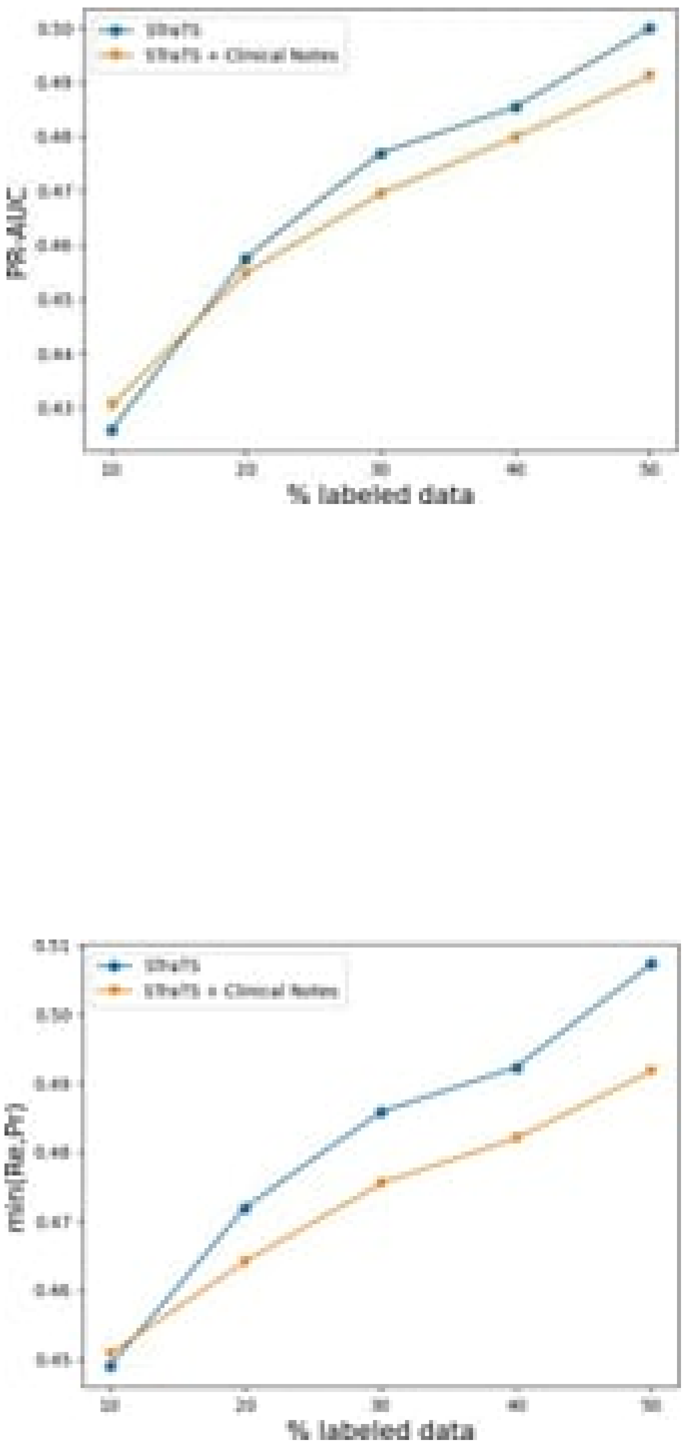
C. STraTS SMALL

I trained STraTS regression and classification models using a smaller subset of patients to create a more balanced dataset. This smaller set included 5,288 septic patients and 10,555 non-septic patients, resulting in a dataset where 33.4% of the patients had a positive sepsis diagnosis. The data was divided into training, validation, and test sets using a 64:16:20 split at the patient level. Table 8 details the number of septic and non-septic patients and ICU stays in this smaller dataset. Figure 3 illustrates the sepsis prediction performance of STraTS models trained on the smaller dataset (STraTS small) compared to the full dataset (STraTS large) and the model incorporating clinical notes (STraTS + clinical notes) from the MIMIC-IV dataset, evaluated across various percentages of labeled data and averaged over 10 runs. As shown in Figure 3, STraTS small exhibits lower ROC-AUC across all labeled data percentages but achieves the highest PR-AUC and min(Re, Pr) values compared to STraTS and STraTS + clinical notes, which were trained on the full dataset with a larger amount of patient data.

Data	Non-septic patients	Septic patients	Non-septic ICU stays	Septic ICU stays
Train	4261	2133	10165	6394
Validation	1071	528	2479	1599
Test	1350	649	3199	1999

Bảng C.1: Number of septic/non-septic patients/ICU stays in train/validation/test data in the smaller dataset.





Hình C.1: Sepsis prediction performance on MIMIC-IV dataset for different percentages of labeled data averaged over 10 runs

D. SEPSIS CHECK COMPONENTS AND VARIABLES

Tables 10 and 11 list the components and their corresponding variable names used for identifying suspected infections and calculating SOFA scores.

SOFA Component	Variable Name
Nervous System	Glasgow Coma Scale (GCS_eye, GCS_verbal, GCS_motor)
Cardiovascular	Mean Arterial Pressure (DBP, SBP)
Administration of Vasopressors	Dopamine, Dobutamine, Epinephrine, Norepinephrine
Respiratory System	FiO2 [kPa] (FiO2)
Mechanical Ventilation	Mechanical ventilation
Coagulation	Platelet Count [x10^3µl] (Platelet Count)
Liver	Bilirubin [mg/dl] (Bilirubin (Total))
Renal	Creatinine [mg/dl] (Creatinine Urine)
	Urine [ml/day] (Urine)

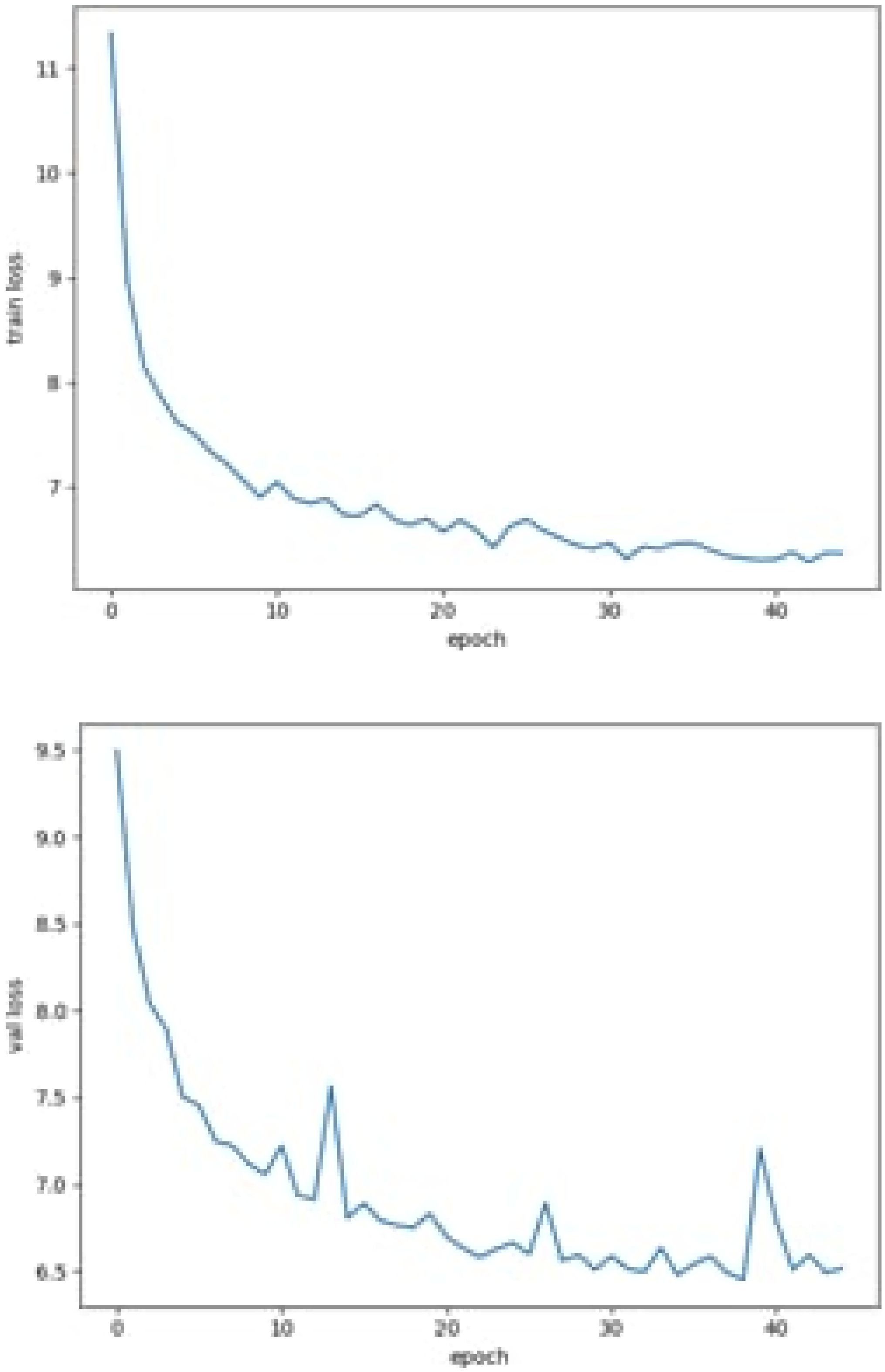
Bảng D.1: Components and corresponding variable names for SOFA.

Suspected Infection Component	Variable Name
Time of Blood Cultures	Blood Cultures
Time of Antibiotics	Antibiotics

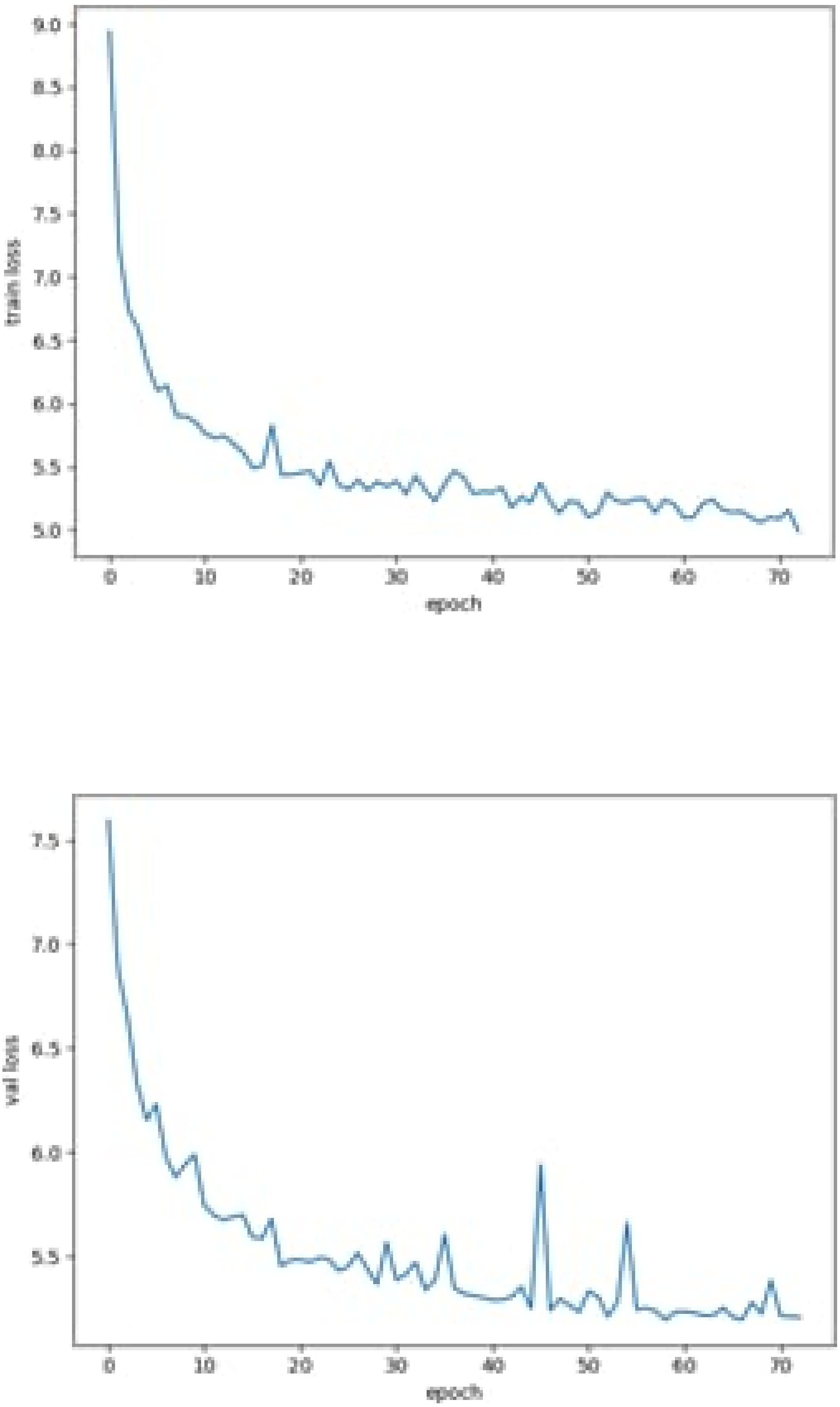
Bảng D.2: Components and corresponding variable names for suspected infection.

E. TRAIN/VALID LOSS

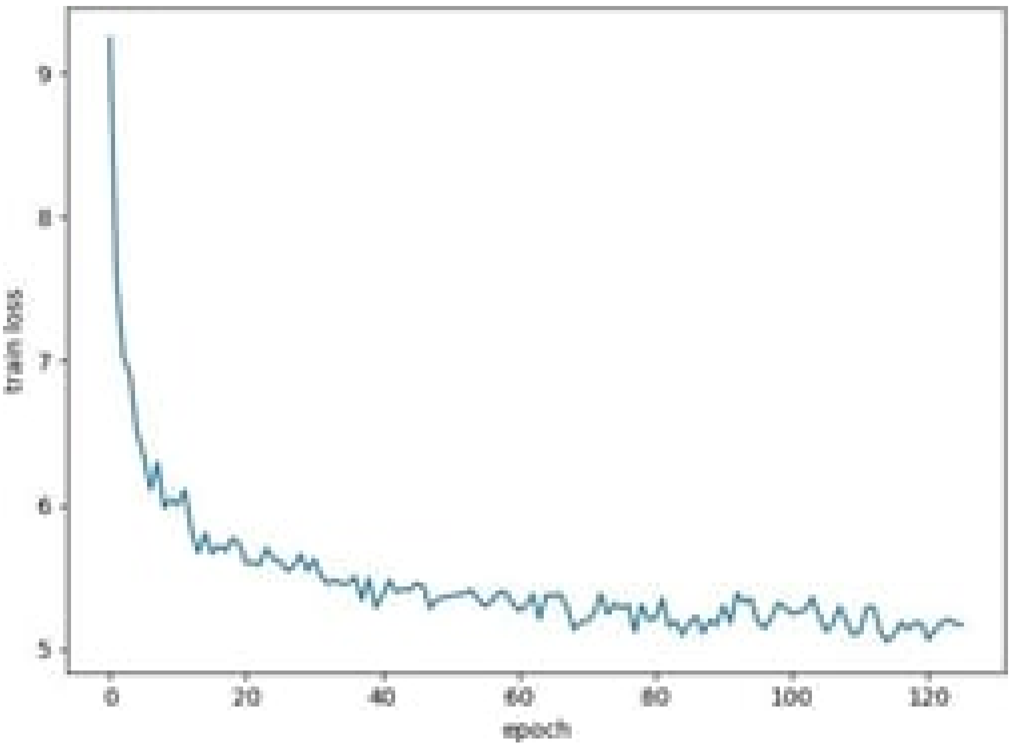
Figure 4, 5, 6 shows train adn validation loss over epochs for STraTS small, STraTS large, and STraTS + Text.

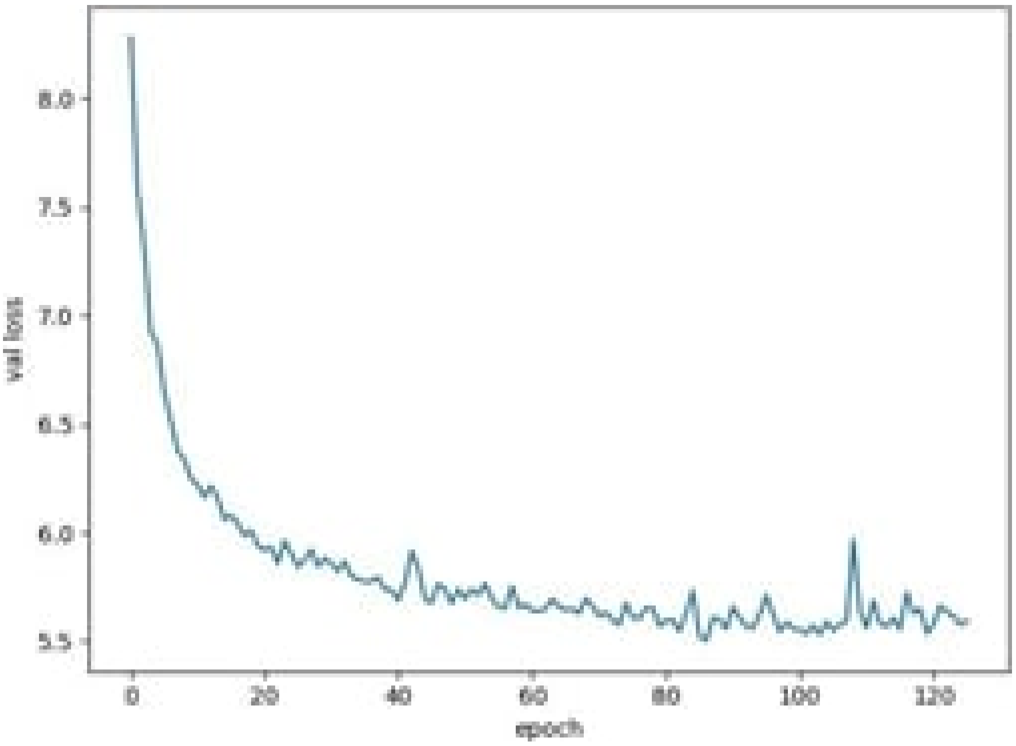


Hình E.1: Train and validation loss over epochs during forecatsing for STraTS small.



Hình E.2: Train and validation loss over epochs during forecatsing for STraTS large.





Hình E.3: Train and validation loss over epochs during forecatsing for STraTS text.