Early Sepsis Prediction: A Rule-based Sepsis Check Supported by Data-driven transformer-based Model

Jinghua Xu*
Heidelberg University
jinghua.xu@stud.uni-heidelberg.de

Natalia Minakova*
Heidelberg University
mina@cl.uni-heidelberg.de

Pablo Ortega Sanchez*
Heidelberg University
ort.san.pablo@gmail.com

Abstract

Sepsis is a serious complication of an infection. Without quick treatment it can lead to organ failure and death. Thus, early detection and treatment of sepsis have the potential to save a significant amount of lives. Yet, their effectiveness often relies on awareness and acceptance of said procedures. In this work, we implement sepsis check based on a widely accepted guideline for sepsis recognition (Sepsis-3). Our implementation achieved F-score as high as 0.874. In addition to implementing the ruled-based approach to early sepsis detection, we use an existing data-driven transformer-based STraTS model (Tipirneni and Reddy, 2021) for time-series forecasting to support sepsis check and directly predicting sepsis label using 24-hour patient data in a fully data-driven setup. Additionally, we attempt to improve mono-modal STraTS model by integrating a clinical text embedding module to enable multi-modal learning. Both the original STraTS model and our refined STraTS+Text model perform good in both forecasting (masked MSE at approx. 5.24) and classification task (ROC-AUC at approx. 0.89).

1 Introduction

Sepsis occurs when the body's immune response to an infection becomes dysregulated and triggers systemic inflammation. It is a leading cause of death in the Intensive Care Units (ICU). Early detection of sepsis is crucial for patient survival (Rudd et al., 2020). In an effort to identify septic patients from their clinical data, Singer et al. (2016) and Reyna et al. (2019) present slightly different approaches. These rule-based guidelines revolve around identifying suspected infections and a clinical criterion for life-threatening organ dysfunction. The guidelines by Singer et al. (2016) were developed as an in-hospital tool to determine the state and condition of a patient. The implementation of these guidelines can be applied to already observed data, but more importantly, could be applied to forecast time-series values based on observed data. This would potentially allow us to identify the development of sepsis earlier and maybe even improve the chances of preventing sepsis. Therefore, we implement a rule-based sepsis check based on a widely accepted guideline for sepsis recognition (Sepsis-3) to enable early sepsis prediction.

In addition to implementing a rule-based sepsis check, we use and refine an existing Self-supervised Transformer for Time-Series (STraTS) model (Tipirneni and Reddy, 2021) for time-series forecasting and 24-hour sepsis prediction. We use the STraTS regression model to forecast time-series values following each observation window to support sepsis check, and the STraTS classification model to predict sepsis label using 24-hour patient data in a fully data-driven setup. On the basis of the original STraTS architecture, which can only take continuous physiological features as its input, we integrate a clinical text embedding module based on Clinical BERT (Alsentzer et al., 2019) to encode 1.4 M clinical notes associated with patients in our data from MIMIC-III. Both models (STraTS and STraTS+Text) show good performance in both forecasting and classification tasks, achieving masked MSE (Mean Squared Error) approximately at 5.24 and ROC-AUC approximately at 0.89.

Our rule-based sepsis check achieved an F-score as high as 0.874 without using features values predicted by the STraTS forecasting model. While introducing predicted values produced by the STraTS forecasting model did not improve the performance of the rule-based sepsis check, the STraTS forecasting predictions helped with the problem of data sparsity and enabled the rule-based sepsis check to identify septic patients whose clinical data alone would not be sufficient to correctly classify them.

*Authors arranged by alphabetical order, order does not indicate significance of contribution by each author.

We release our code at: github.com/JINHXu/Research-Module-Early-Sepsis-Prediction.

2 Related Work

Torio and Andrews (2013) found that in 2011, sepsis was the most expensive condition treated in U.S. hospitals and accounted for 5.2% of the total hospitalization costs. In a more global context, Rudd et al. (2020) found that in 2017 sepsis resulted in 19.7% of all global deaths, with the most cases in low- or middle income countries. While guidelines for identifying and treating septic patients do exist, like the one proposed by Singer et al. (2016), Kissoon (2014) argue that many factors complicate their on-site implementation. Depending on the environment and resources, these factors include critical staff shortages, failure to identify sepsis and lack of availability of laboratory tests. Previous hypothesis regarding the development and diagnosis of sepsis heavily relied on the Systemic Inflammatory Response-Syndrome (SIRS) (Bone et al., 1992). More recently, the reliability of SIRS has been found to be not sufficient which prompted the development of the Sepsis-related Organ Failure Assessment (SOFA) (Vincent et al., 1996).

With the growing use of electronic health record (EHR) systems, clinical time-series data and digital clinical notes have been used to enable machine learninng models for prediction tasks in the medical domain such as sepsis (Wang et al., 2022) and mortality prediction (Tipirneni and Reddy, 2021). Earlier studies applied linear dynamical systems (LDS) (Liu et al., 2013) and Gaussian process (GP) (Liu and Hauskrecht, 2015) to model time-series data in the clinical domain. Meanwhile models such as BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019) pretrained on text data in the medical domain have been proposed to model articles and doctor-written notes in the medical domain.

Moving past the phase of simply modeling data, more recent studies seek to solve the problem in specific predictive tasks. Wang et al. (2022) proposed a multi-modal learning solution to early sepsis prediction by integrating clinical notes and continuous physiological features to construct a transformer-based binary classification model. While the model in Wang et al. (2022) achieved good performance on MIMIC-III (Johnson et al., 2016) and eICU-CRD (Pollard et al., 2018) datasets to predict a binary label indicating patient septic state based on observed data following admission into ICU, it is not capable of forecasting time-series values. Tipirneni and Reddy (2021) proposed a two-step transformer-based approach to mortality prediction, including a regression model to forecast time-series values and a classification model to predict patient septic label.

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. Thus, one of the key indicators for septic patients is a suspected infection. With only data at hand, a suspected infection is identified by orders for blood cultures and antibiotics. The Singer et al. (2016) guideline only accounts for a specific time period in which antibiotics and cultures are ordered. If antibiotics were administered first, blood cultures need to be taken within 24 hours and if blood cultures were taken first, antibiotics need to be administered within 72 hours. In the implementation, this time period is called suspicion window. In contrast, Reyna et al. (2019) require antibiotics to be administered for at least 72 consecutive hours to be considered as a suspected infection. If that is the case, the first administration of antibiotics is compared to blood culture orders identically to Singer et al. (2016). This distinction has a rather great impact on the implementation, as the sparsity of available data and therefore the ability to identify consecutive administrations of antibiotics is a non-trivial challenge.

3.2 Criterion for life-threatening organ dysfunction

Another key indicator for septic patients is life-threatening organ dysfunction. There are several criterions that can be employed to identify life-threatening organ dysfunction, however, both Singer et al. (2016) and Reyna et al. (2019) suggest the Sequential [Sepsis-related] Organ Function Assessment (SOFA) (Vincent et al., 1996), which takes into account a variety of clinical and laboratory variables. Following Singer et al. (2016), a patient's SOFA score should be computed for each time step – which in this case means per hour. The time of SOFA is then the time at which a patient

gets a SOFA score of two or higher, considering the initial value to be zero if no organ dysfunction is known beforehand. Reyna et al. (2019) consider the time of SOFA to be the time at which an increase of two in comparison to the last 24 hours happens. If this time of SOFA is at most 48 hours before or 24 hours after a suspected infection, the patient developed sepsis according to Singer et al. (2016). Reyna et al. (2019) are more strict and only allow the time of SOFA to be at most 24 hours before and 12 hours after the time of a suspected infection. Additionally, they treat the earlier of the two times as the time of onset of sepsis. The aforementioned time period between suspected infection and time of SOFA is called sepsis window in the implementation.

3.3 Implementation

In order to compute either Singer et al. (2016) or Reyna et al. (2019) checks, the clinical features described in Tables 11 and 10 should be reported in the patient data, with a great importance on antibiotics and blood culture features. Given the nature of the rule-based guidelines, the lack of either the antibiotics or blood culture feature will always result in a negative sepsis label, as these features are directly responsible for suspecting an infection.

3.4 About preprocessing and running the sepsis check

The sepsis check comprises several utility functions to process Tipirneni and Reddy (2021)'s output. Before starting an experiment one needs to decide if 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 of Tipirneni and Reddy (2021) includes a normalization in the form of:

$$normalized = \frac{value - mean}{std} \tag{1}$$

Consequently, numerical features are re-normalized 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 feature for antibiotics is filled depending on what was decided beforehand. Then, the features are cast to their correct type.

3.5 SOFA

Before the sofa scores can be computed, the Glasgow Coma Scale is computed from its components and the mean arterial pressure is estimated using diastolic and systolic blood pressures according to Demers and Wachs (2019) by:

$$DBP + \frac{SBP - DBP}{3} \tag{2}$$

Now the sofa score is computed for each hour, after which a 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 to be the time of suspicion, which is then compared to the time of SOFA under the constraints of the sepsis window. As already mentioned, this part of the sepsis check is most critical. If either blood cultures or antibiotics are not reported in the patient data, the patient did not develop sepsis according to the guidelines. During the evaluation of the first experiments, it became clear that, between the suspicion window, strategy and fill procedure, the number one reason for erroneous classifications of patients was the lack of a time of suspicion. At first, the suspicion window was suspected to be the culprit. Increasing the suspicion window to up to ten days increased the suspected infections, however, now, they were suspected too late and missed the sepsis window. This is more serious for Reyna et al. (2019), because antibiotics need to be administered 72 consecutive hours. If a hospital stay is less than 72 hours, there can be no sepsis. And 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 tackle this problem, two more strategies were implemented. While the Sepsis-3 strategy uses the first time of blood cultures, and the first time of antibiotics and the supplied suspicion window to compute the time of suspicion, the 'catchsus' strategy takes into consideration all times blood cultures were taken and all times antibiotics were administered. It then checks if any of the possible combinations of time of antibiotics and time of

blood cultures fall within the specified suspicion window. For each of the possible combinations that fall within the specified suspicion window, the earlier time is considered the time of suspicion. This can yield multiple times of suspicions, which are then compared to the time of sofa and the sepsis window to generate a sepsis label. The 'grouped' strategy takes this approach even further and expands it onto the time of sofa. As a result, multiple times of suspicion and multiple times of sofa are considered when generating a sepsis label. Unfortunately, even though 'catchsus' and 'grouped' strategies outperformed the standard strategies on unfiltered data – which contains a lot of 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 a better strategy.

3.7 Utilizing Time-Series Forecasting

Next to the obvious benefits of potentially being able to predict the onset of sepsis before the features that would indicate said sepsis are even observed, another advantage of utilizing time-series forecasting is, that the important features can be forecast as well. Unfortunately, in this case, both antibiotics and blood cultures are binary features, whereas Tipirneni and Reddy (2021) is designed to output continuous values. To interpret these continuous values, a threshold was set that assigns everything above or below it a binary label. The method to find said threshold is improvable. So far, a combination of clustering and careful trial and error was used. The experiments that were conducted for this research paper combine observed data and one hour of forecasting output based on that observed data. For each observation window from 20 to 120 hours in steps of four, the resulting concatenation is used as input for the sepsis check.

4 Data

4.1 MIMIC-III

Medical Information Mart for Intensive Care 3 (MIMIC-III) is a large database consisting of patients who stayed in the critical care unit at the Beth Israel Deaconess Medical Center between 2001 and 2012. Johnson et al. (2016) The database consists of twenty six tables. Some examples of the tables include clinical notes, chartevents, admissions and microbiology events. For the full list of tables please refer to Table 4 in Johnson et al. (2016) or to the data exploration part of our code.

4.2 Sepsis Label Annotation

Our project is based on the STraTS architecture (Tipirneni and Reddy, 2021), and as such, we utilized their preprocessing approach to prepare the data in the required format. The authors of Tipirneni and Reddy (2021), however, focus on mortality prediction, which requires a mortality label. In our approach to predict sepsis we need to perform a sepsis check that was introduced earlier in this paper. To identify and label patients with sepsis, ICD9 codes from the diagnosis table have been used. In total, we used 23 codes related to sepsis, as listed in Table 7 in Appendix B. After the data is generated, a sepsis label is assigned to the hospital admission id. Furthermore, we filtered out patients who were admitted with sepsis from our dataset, by parsing the admissions table from MIMIC-III. ICD9 codes are not present in the admission table, therefore, the patients were filtered out based on string matches. These patients have been excluded because the forecasting model cannot benefit by learning from them.

4.3 Our Data

From MIMIC-III dataset, we built our data from 5288 septic patients (9.2%) and 51994 non-septic patients. We split data into train, validation, test by 64: 16: 20 at patient level (table 1).

In addition to the 133 physiological features (see full list in Appendix A), we include 1,407,430 clinical notes from the MIMIC-III dataset in our data. Table 2 shows text statistics on clinical notes associated with patients included in our data.

4.4 Clinical Notes Preprocessing

We preprocess clinical notes following common practice for clinical text cleaning by removing stop words and special characters, and normalising case. Additionaly, in order to avoid potential label leakage in the STraTS classification task, we remove sentences containing "sepsis" or "septic".

5 Models

5.1 STraTS

For both time-series forecasting of physiological feature values for our rule-based sepsis check and

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

Table 1: Number of septic/non-septic patients/ICU stays in train/validation/test data.

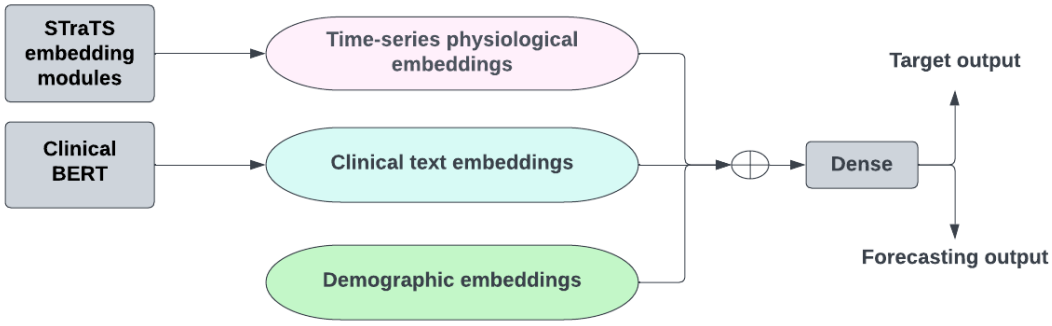

Figure 1: STraTS + Clinical Text Embedding Architecture.

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

Table 2: String length and token counts in clinical notes included in our data.

direct sepsis label prediction, we use the existing Self-supervised Transformer for Time-Series (STraTS) model (Tipirneni and Reddy, 2021). The STraTS model encodes time-series physiological data in an anti-conventional manner: it represents continuous data using a novel Continuous Value Embedding technique to avoid discretizing data (e.g. aggregation, imputation). Each observation is represented as a triplet: observed time, variable name, and variable value. After initial triplet embedding, the embeddings go through a transformer component with multi-head attention layers to enable learning contextual triplet embeddings, followed by a fusion self-attention module to complete embedding time-series data.

The STraTS architecture supports both a regression model for time-series forecasting and a binary classification model to predict a state label (e.g. mortality, sepsis). The STraTS regression model was originally designed to deal with limited availability of labeled data in the medical domain. It is used during forecasting as an auxiliary proxy task to optimize performance for the classification model. In our case, we fully use both models to obtain forecasting results to support our rule-based sepsis check, and the classification model for 24-hour septic state prediction in a fully data-driven setup.

5.2 STraTS + Clinical Text Embedding

On the basis of the original STraTS model, which encodes only physiological features as its input for forecasting and classification, we additionally add a clinical text embedding module to the architecture. We embed clinical notes using clinical BERT (Alsentzer et al., 2019) to obtain text features, and align text features side by side with time-series embedding of physiological features and demographic features for concatenation, and pass through a dense layer to generate outputs for both forecasting and classification task. Figure 1 shows the architecture of the refined STraTS with clinical text embedding.

6 Results & Discussion

6.1 STraTS Forecasting

We train STraTS regression models with and without clinical text embeddings to forecast physiological feature values in the two hours following the observation windows, defined as $\{min(0, x - 24), x) | 20 \leq x \leq 124, x\%4 = 0\}$. We obtain predictions of both regression models on test data to support rule-based sepsis check. We use masked MSE (mean squared error) for evaluation, where the binary mask indicates if a true value was observed in data.

Table 3 shows masked MSE (mean squared error) on test and validation data for STraTS and STraTS + Text regression models. It can be seen from the table that both regression models show similar performance on test data, while the original STraTS without introducing clinical text embeddings had a better MSE on validation data.

Model	Test	Validation
STraTS	5.2455	5.2048
STraTS + Text	5.2493	5.5922

Table 3: Masked MSE (mean squared error) on test and validation data for STraTS and STraTS + Text models.

6.2 STraTS Classification

We use 24 hour data (after admission to ICU) to train STraTS classification model with and without clinical text embeddings using random sample of 10, 20, 30, 40, 50 % labelled data to predict septic state for each ICU stay. We repeat each experiment 10 times with different randomly sampled data from train and validation sets.

In the binary classification task, we use three metrics to evaluate model performance on sepsis prediction:

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

Figure 2 shows ROC-AUC, PR-AUC, and min(Re, Pr) of both models evaluated on the test dataset. It can be seen from the charts that STraTS + Clinical Text only slightly outperforms STraTS when the percentage of labelled data is lower ($\leq 30\%$) in terms of ROC-AUC score. The STraTS model in general has higher PR-AUC and min(Re, Pr) except when only 10% of labelled data is available. STraTS + Clinical Text only shows slightly more advance performance when it in a low-amount labelled data setup, whereas STraTS has better performance with more available labelled data.

6.3 Sepsis Check results

Since the experiments that utilize time-series forecasting only add one hour of unobserved data, the results in table 5 are not as significant as we intended. However, it should be noted that the additional data led to more correctly identified true positives in all sets of experiments and even for patients where the original observed data does not contain both antibiotics and blood culture features. This means that STraTS may enable the sepsis check to overcome the problem of data sparsity. Expanding the STraTS output to multiple hours of predictions seems to be a promising direction for future research.

7 Conclusion

We implement a rule-based sepsis check and found that the sparseness of patient data is greatly impacting the potential performance. Our rule-based check achieved good performance by itself, and we find in our experiments that by introducing values forecast by the STraTS regression model in its current form did not improve sepsis check. Additionally, our approach can possibly help to tackle the many complicating real-world factors that can complicate sepsis treatment that are outlined in Kissoon (2014).

On top of the rule-based sepsis check, we use and refine an existing STraTS model (Tipirneni and Reddy, 2021) for time-series forecasting to support sepsis check and directly predicting sepsis label using 24-hour patient data in a fully data-driven setup. We attempted to improve the STraTS model by integrating a clinical text embedding module based on Clinical BERT to enable multi-modal learning by taking both text and physiological features into account. Both STraTS and STraTS+Text regression models achieved good performance on time-series value forecasting. The classification models also showed high ROC-AUC score in the task of sepsis prediction using features selected in our study. However, the current STraTS+Text model does not outperform the original STraTS model in our experiments. For future work, we intend to continue improving the STraTS+Text model architecture, mainly the text embedding module, to enable better representation. Additionally, in order to better support the rule-based sepsis check, we plan to further improve the STraTS architecture to enable extended forecasting window based on a short and limited forecasting observation window.

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

Table 4: Sepsis prediction performance on MIMIC-III dataset. The results show mean and standard deviation of the metrics after repeating the experiment 10 times by sampling 50% labeled data each time.

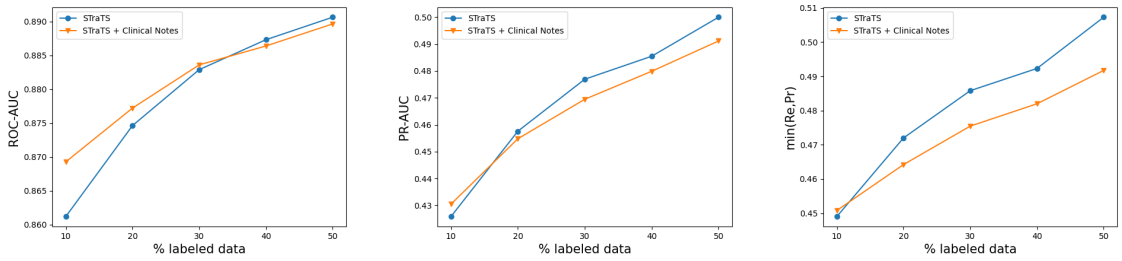

Figure 2: Sepsis prediction performance on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs.

experiment	F1 score	true positives
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**

Table 5: 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.

Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

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.

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.

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

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Nature*.

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

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.

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

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.

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.

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.

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.

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.

Sindhu Tipirneni and Chandan K. Reddy. 2021. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series.

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

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.

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

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
Eoisinophils
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 Regular

Insulin 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-III

The following section contains the sepsis codes from the D_ICD_DIAGNOSES table that were used to determined the positive label for the data. The codes can be found in Table 7.

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. influenae septicemia
03842	E coli septicemia
03843	Pseudomonas septicemia
03844	Serratia septicemia
03849	Gram-neg septicemia NEC
0388	Septicemia NEC
0389	epticemia 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

Table 7: ICD9 Codes for sepsis

C STraTS small

We also trained STraTS regression and classification models with data of a smaller set of patients. We used the same 5288 septic patient data and 10555 non-septic patients, resulting in a more balanced dataset (33.4% positive class at patient level). We split data into train/validation/test at patient level also by 64:16:20. Table 8 shows the number of septic/non-septic patients/ICU stays in the smaller dataset.

Figure 3 shows sepsis prediction performance of STraTS small, STraTS large and STraTS + clinical notes on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs. As shown in 3, STraTS small shows lower ROC-AUC with all percentages of labelled data, whereas it has the highest PR-AUC and min(Re,Pr) compared to STraTS and STraTS + clinical notes, which were trained on the full dataset with more patient data.

D Sepsis check components and variables

Table 11 and table 10 show the components and corresponding variable names within the data for identifying suspected infections and calculating SOFA.

E Train/valid loss

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

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

Table 8: Number of septic/non-septic patients/ICU stays in train/validation/test data in the smaller dataset.

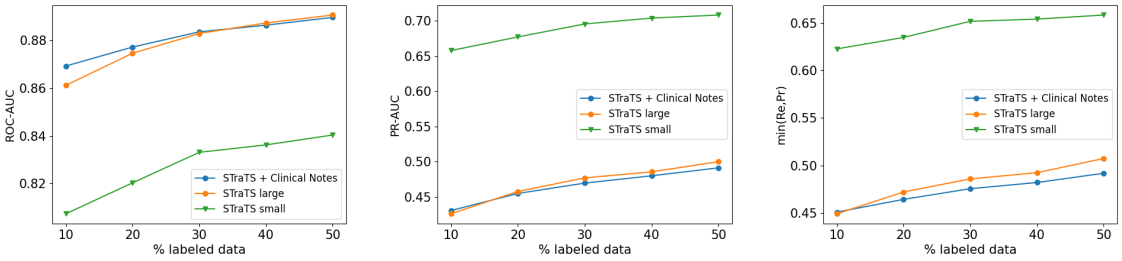

Figure 3: Sepsis prediction performance on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs.

Model	ROC-AUC	PR-AUC	min(Re,Pr)
STraTS small	0.840 ± 0.003	0.708 ± 0.005	0.658 ± 0.006
STraTS large	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

Table 9: Sepsis prediction performance on MIMIC-III dataset. The results show mean and standard deviation of the metrics after repeating the experiment 10 times by sampling 50% labeled data each time.

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 [x$10^3 \mu$l]	Platelet Count
Liver	Bilirubin [mg/dl]	Bilirubin (Total)
Renal	Creatinine [mg/dl]	Creatinine Urine
	Urine [ml/day]	Urine

Table 10: Components and corresponding variable names for SOFA

Suspected Infection	
component	**variable name**
time of blood cultures	Blood Cultures
time of antibiotics	Antibiotics

Table 11: Components and corresponding variable names for suspected infection

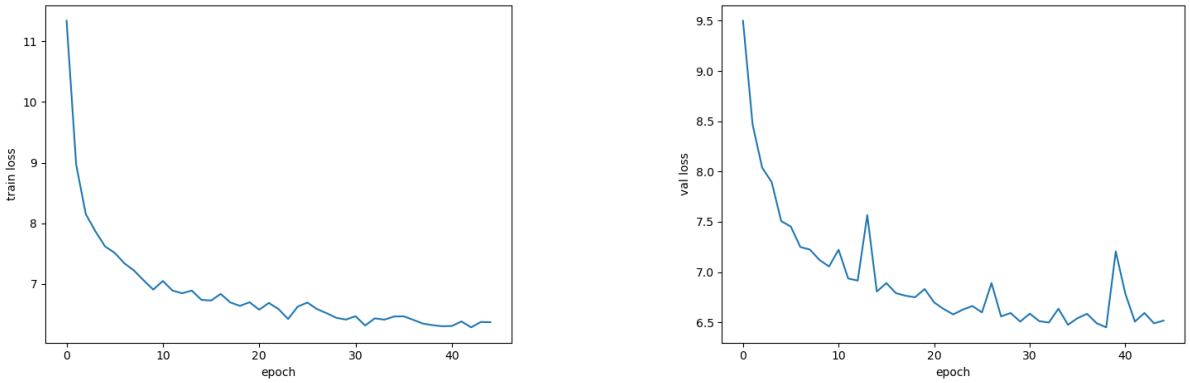

Figure 4: Train and validation loss over epochs during forecatsing for STraTS small.

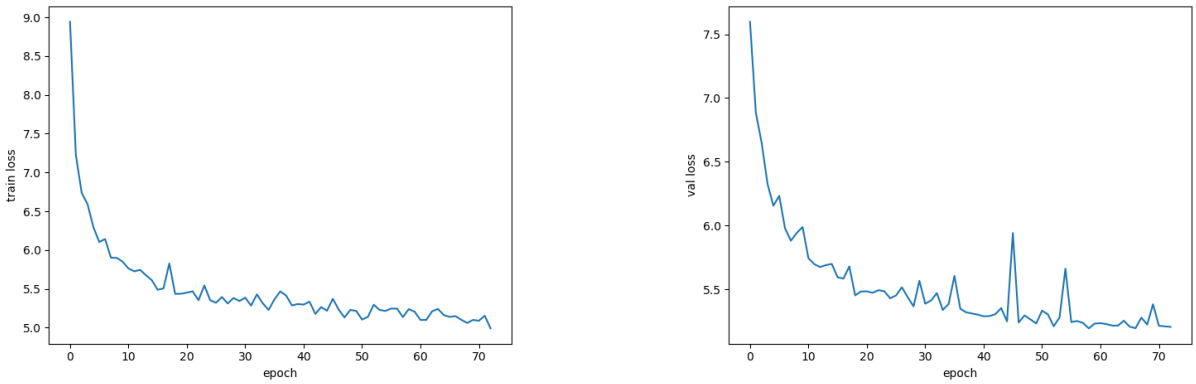

Figure 5: Train and validation loss over epochs during forecatsing for STraTS large.

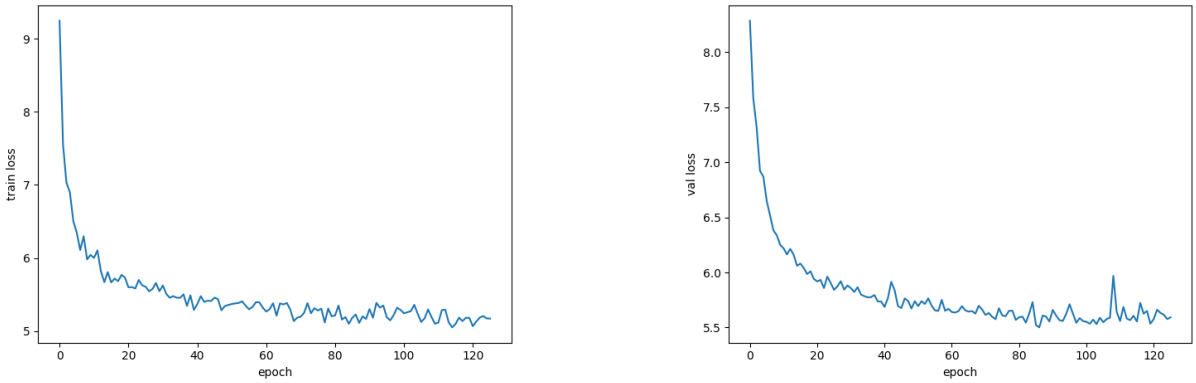

Figure 6: Train and validation loss over epochs during forecatsing for STraTS text.