Multimodal Transformers for Clinical Time Series Forecasting and Early Sepsis Prediction

Jinghua Xu
Heidelberg University, Germany
xu@cl.uni-heidelberg.de

Michael Staniek
Heidelberg University, Germany
staniek@cl.uni-heidelberg.de

Abstract

Sepsis is a leading cause of death in Intensive Care Units (ICU). Early detection of sepsis is crucial to patient survival. Existing works in the clinical domain focus mainly on directly predicting a ground truth label that is the outcome of a medical syndrome or condition such as sepsis. In this work, we primarily focus on clinical time series forecasting as a means to solve downstream predictive tasks intermediately. We base our work on a strong monomodal baseline and propose multimodal transformers using set functions via fusing both physiological features and texts in electronic health record (EHR) data. Furthermore, we propose hierarchical transformers to effectively represent clinical document time series via attention mechanism and continuous time encoding. Our multimodal models significantly outperform baseline on MIMIC-III data by notable gaps. Our ablation analysis show that our atomic approaches to multimodal fusion and hierarchical transformers for document series embedding are effective in forecasting. We further fine-tune the forecasting models with labelled data and found some of the multimodal models consistently outperforming baseline on downstream sepsis prediction task.

1 Introduction

Sepsis is a serious complication of an infection, accounting for approximately 19.7% of all global deaths (Rudd et al., 2020). In 2017, World Health Organization declared that improving the prevention, recognition, and treatment of sepsis as a global health priority (WHO, 2020). Seymour et al. (2017) and Liu et al. (2017) suggest an increase in the adjusted mortality of septic patients with delayed antibiotic administration. With patients suffering from septic shock, Kumar et al. (2006) found an 3.6–9.9% hourly increase in mortality when treatment is delayed. Early Detection of Sepsis is critical to improve patient outcome.

With the emerging abundance of clinical electronic health record (EHR) data, multimodal patient data present both challenges and opportunities to forecasting and predictive tasks in the clinical domain. On the one hand, multimodal representation learning is a complex problem that requires proper handling of information from multiple sources (Tsai et al., 2018). On the other hand, data from various sources enrich information available to models, which enables more robust prediction (Baltrušaitis et al., 2018). Fusing multiple modalities such as laboratory measurements, clinical texts, medications, and procedures have shown improved performance on predicting inpatient mortality, length of stay, and 30-day readmission (Rajkomar et al., 2018).

A further challenge in learning from clinical EHR datasets lies with data missingness and irregularity. The available observations for each patient may vary based on patient's condition, i.e. the set of observed clinical variables for each patient can differ from one another. Additionally, clinical measurements are often not taken at regular time intervals - the measurements may occur sporadically in time depending on the underlying conditions of the patient. Previous works such as Wang et al. (2022) simply aggregate data into hourly bins to circumvent data missingness, irregularity and sporadicity. However, this introduces noises and suppresses information to indicate patient condition through the actual availability of clinical measurements. To tackle the issue, Tipirneni and Reddy (2022) implements "Triplet Embedding" based on Set Functions proposed in Horn et al. (2020) to represent each clinical observation for each patient at each time discretely to avoid data imputation/aggregation of any form. While Tipirneni and Reddy (2022) achieves excellent performance on prediction tasks against several strong baselines, it disregards information potentially contained in clinical notes associated with each patient

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 100–108
May 4, 2025 ©2025 Association for Computational Linguistics

Paper	Multi-modal	Set Function	Time Encoding	Forecasting
Horn et al. (2020)	✗	✓	Sinusoidal Encoding	✗
Wang et al. (2022)	✓	✗	✗	✗
Lyu et al. (2022)	✓	✗	Sinusoidal Encoding	✗
Tipirneni and Reddy (2022)	✗	✓	Learnable Embedding	✓
Lee et al. (2023)	✓	✓	Linear Projection	✗
Proposed Models	✓	✓	Leanrable Embedding	✓

Table 1: Tabular comparison of proposed models and related works closely referred to.

record in EHR data.

With majority existing works in the clinical domain approach predictive tasks directly by predicting a ground truth label as the outcome of observed patient conditions (Lee et al., 2023; Tipirneni and Reddy, 2022; Wang et al., 2022; Lyu et al., 2022), Xu et al. (2023) proposed to focus on forecasting, and implemented a rule-based sepsis check for Sepsis prediction that depends on model forecasts. We follow this practice and primarily seek to build models for time series forecasting (cause prediction), as an intermediate means to eventually predict sepsis and potentially other medical syndrome instead of predicting an outcome directly.

To address various limitations with existing works, we build upon a strong monomodal baseline model (Tipirneni and Reddy, 2022) and propose multimodal transformers primarily for clinical time series forecasting that 1) incorporates information from both physiological time series data and clinical notes via effective multimodal fusion 2) utilizes set functions to avoid data aggregation and imputation. The forecasting models produce predictions of the clinical variable values in a two-hour forecasting window following corresponding observation windows of varying lengths, to support ruled-based implementations (e.g. Xu et al., 2023) that rely on predicted values of specific clinical variables. Meanwhile, the forecasting models are fine-tunable with labelled data for downstream prediction tasks such as sepsis prediction. We additionally propose a hierarchical transformer to effectively represent clinical notes that naturally form document time series within observation windows by integrating time embeddings of note records, and accounting for the interactions between notes in time order via attention mechanism. We conduct comprehensive experiments and ablation analysis to showcase that our proposed models and the atomic modules are effectively robust, improving forecasting performance from baseline significantly.

We summarise the main contributions of our work as follows:

- We propose a multimodal learning framework for patient data in EHR datasets that effectively incorporates information from both physiological features and associated clinical notes.
- We propose a specialized hierarchical transformer to effectively represent clinical document time series that accounts for the interactions between individual clinical notes via attention and brings cross-modal time awareness to the entire model through consistent time encoding.
- Our clinical time series forecasting models approach predictive tasks in the clinical domain from a cause-prediction perspective. It provides flexibility in two dimensions: 1) the forecast values can be used for prediction of multiple medical syndromes and conditions with rule-based implementations (e.g. sepsis check based on Sepsis-3 definition (Reyna et al., 2020; Seymour et al., 2016; Singer et al., 2016)) 2) the forecasting models can be fine-tuned for arbitrary downstream prediction tasks with correspondingly labelled data in a fully data-driven setup. Additionally, the intermediate results produced by forecasting models are also directly interpretable by clinical practitioners as pointed out in previous work.

We release our code at github.com/JINHXu/clinical-multimodal-transformers.

2 Related Work

Clinical time series data are inherently sequential, making common sequence modelling methods (RNNs, transformers, etc.) suitable. Early works use classic models such as Gaussian Process (GP) (Liu et al., 2013, 2017; Lu et al., 2008; Li and Marlin, 2016) and linear dynamical systems

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

(LDS) (Liu and Hauskrecht, 2015) to model irregular clinical time series. Later works then employ RNN-based models given the sequential nature of time series data. Baytas et al. (2017), for instance, modified LSTM to fit hidden cell states to irregular time slots (T-LSTM). Che et al. (2018), on the other hand, modified the GRU cell which decays inputs to global means and hidden states through unobserved time intervals (GRU-D). The problem with classic models such as Gaussian Process are their sensitivity to choice of covariance and mean functions, while RNNs process long sequences (resulted by irregularity) sequentially with inability to parallel computation thus leading to long runtime.

More recent works employ transformer-based methods for clinical time series modeling. Wang et al. (2022), for instance, passes multivariate time series embeddings first through a block of transformer encoders to capture contextual information of the sequences, then followed by a dense interpolation layer to obtain a concise representation of transformer outputs. Tipirneni and Reddy (2022) also uses multi-head attention to obtain contextual embeddings through transformer, it then passes these embeddings to a self-attention layer to capture the context within each observation. Horn et al. (2020) uses attention-based aggregation to compute embeddings of set elements independently from other elements in order to reduce runtime complexity to linear from the original transformer (Vaswani et al., 2017), which accounts for dependency between such elements leading runtime and space complexity of $O(N^2)$. It is worth noting that, in this case, Horn et al. (2020) compromises on accuracy by disregarding such dependency for lower space and runtime demand, while Tipirneni and Reddy (2022) uses a transformer block similar to Vaswani et al. (2017) to guarantee model performance with the expense of computing power and time.

Most works in clinical machine learning focus on predictive tasks. Tipirneni and Reddy (2022) proposes an encoder-only transformer model for direct mortality prediction as the target task, while its intermedial proxy model could be used for time series forecasting. Staniek et al. (2024) proposes encoder-decoder long-term clinical time series forecasting models to predict outcome via predicting the cause of syndromes intermediately. These forecasting models, however, are monomodal models that learn from data of single modality, disregarding potential information delivered through associated clinical notes in EHR datasets.

Multimodal learning is a common practice to address various tasks in the clinical doamin due to the various modalities of data in EHR datasets. Wang et al. (2022) uses concatenation to integrate multimodal patient data on physiological features and clinical texts. Later works in the clinical domain such as (Lyu et al., 2022) additionally applies a multimodal fusion encoder after concatenation of two modalities, in order to map them into a universal space before feeding the embeddings into a transformer. More recent works in the clinical domain employ attention-based fusion methods to represent multimodal patient data. (Lee et al., 2023), for instance, modified attention bottlenecks (Nagrani et al., 2021) from an audio-vision task to learn multi-modal EHR data (EHR time-series, EHR texts, EHR images) for mortality, vasopressor need, and intubation need prediction tasks.

Table 1 presents a tabular review of related works we closely refer to in this work. We seek to tackle the limitations in previous works, and define our primary task as time series forecasting on multimodal patient data. In the following sections, we further lay out the implementation specifics of our models and methods to overcome limitations in existing works.

3 Data

3.1 MIMIC-III

MIMIC-III (Medical Information Mart for Intensive Care 3) is a large database consisting of ICU (Critical Care Unit) patient records at the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). The entire MIMIC-

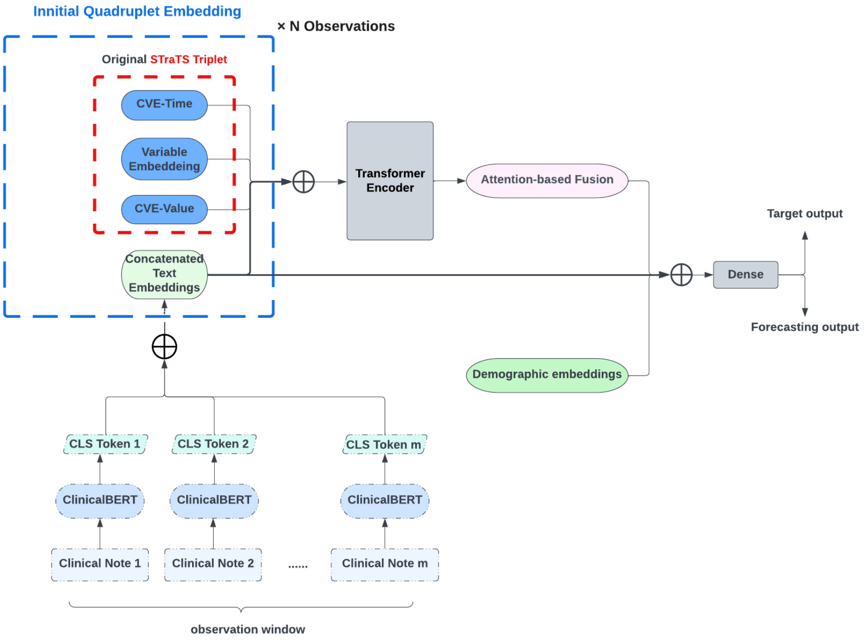

Figure 1: Multimodal STraTS-Q-M - ClinicalBERT

III database stores 61,532 ICU stays among 58, 976 hospital admissions from 46,520 patients. The database is composed of 26 tables including clinical notes, chartevents, admissions and microbiology events and etc.

3.2 Our Data

We use annotated data with septic patients labelled based on 23 ICD-9 codes and the Sepsis-3 definitions (Reyna et al., 2020; Seymour et al., 2016; Singer et al., 2016). Patients admitted with sepsis were excluded from experiment data as they may mislead model in fine-tuning stage for sepsis prediction. 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. We extract 133 physiological features (record time, feature value) and two demographic features (age and gender) for each admission from MIMIC-III, and include 1,407,430 clinical notes associated with patient records.

3.3 Clinical Note Preprocessing

Prior practices (Wang et al., 2022) conduct stop word and special character removal, case normalisation on clinical notes as text cleaning steps before feeding to a language model such as ClinicalBERT (Alsentzer et al., 2019). We argue that for a contextual language model pretrained on clinical notes without the above mentioned text preprocessing steps, the above cleaning procedures are unnecessary and potentially harmful. As pointed out in Khattak et al. (2019), case normalisation can introduce noise to clinical texts. For instance, by lowercasing the medical condition term ADD (attention deficit disorder), it converts to a verb “add" that leads to ambiguity. Thus we reserve the original clinical notes for ClinicalBERT-based text embedding modules in our models to generate document-level embeddings. With the GloVe-based models, we remove special characters and stop words to reduce noise and improve training efficiency, as necessary text cleaning steps.

4 Methods

4.1 Baseline STraTS

We base our work on a strong baseline model STraTS (Tipirneni and Reddy, 2022), which takes multivariate clinical variables as its monomodal input, encoded by a learnable continuous value embedding module and feature map. STraTS uses set functions to represent clinical time series as triplets to avoid data imputation and aggregation. The encoded triplets are then fed into transformer blocks and a self-attention module to account for the interactions across data instances and triplets within

Model	**Parameters**	**Best Epoch**	**Test**	**Validation**
STraTS *(baseline)*	71,070	71	5.2631	5.2089
STraTS - $ClinicalBERT_{CLS\_emb}$ - *base*	10,230,720	104	5.1771	5.0803
STraTS - $ClinicalBERT_{CLS\_emb}$ - *large*	33,920,820	126	5.2014	5.1226
STraTS-Q - $ClinicalBERT_{CLS\_emb}$	61,140,480	124	5.1742	5.1198
STraTS-Q-M - $ClinicalBERT_{CLS\_emb}$	33,920,820	105	**5.1650**	**5.1152**
STraTS-Q-M - $ClinicalBERT_{avg\_emb}$	33,920,820	101	5.2789	5.1950
STraTS - $GloVe$ - *base*	92,820	86	5.2781	5.1875
STraTS - $GloVe$ - *large*	112,920	109	5.3695	5.1707
STraTS-Q-M - Hierarchical Transformer - *base*	11,605,610	110	5.3312	5.2295
STraTS-Q-M - Hierarchical Transformer - *large*	48,216,860	103	5.2584	5.1836
STraTS-Q-M - Hierarchical Transformer - *large*[1]	48,216,860	147	**5.1535**	**5.0038**

1 Learning rate reduced to 0.0001 after 80 epochs

Table 3: Masked MSE (mean squared error) on test and validation data for each model. (patience = 15, parameters refers to trainable parameters)

Model	**p-value**
STraTS - $ClinicalBERT_{CLS\_emb}$ - *base*	0.0
STraTS - $ClinicalBERT_{CLS\_emb}$ - *large*	0.0
STraTS-Q - $ClinicalBERT_{CLS\_emb}$	0.0
STraTS-Q-M - $ClinicalBERT_{CLS\_emb}$	0.0
STraTS-Q-M - Hierarchical Transformer - *large*	0.69
STraTS-Q-M - Hierarchical Transformer - *large*[1]	0.0

Table 4: Randomization test results for proposed models against baseline on forecasting task.

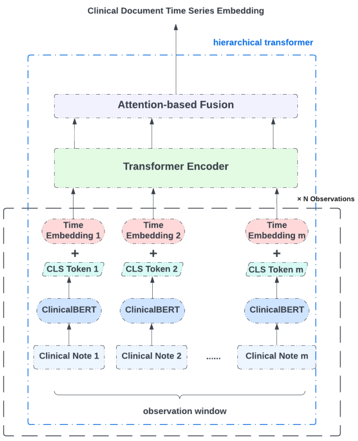

Figure 2: Hierarchical Transformer for Clinical Document Time Series

each observation window.

4.2 Multimodal STraTS-Q-M - ClinicalBERT

On the basis of STraTS, we further include associated clinical notes represented by document-level embeddings obtained through ClinicalBERT. We first obtain initial quadruplet embedding instead of triplets in STraTS as follows:

$$e_i = e_i^f + e_i^v + e_i^t + e_i^T \tag{1}$$

where e_i^f, e_i^v, e_i^t are feature, value, time embeddings originally to form the triplets, along with the associated text embedding e_i^T aligned by observation windows.

The initial quadruplet embeddings are then passed to the following transformer blocks and self-attention module. Eventually, we obtain a fused multimodal representation via concatenating with demographic feature embeddings and ClinicalBERT text embeddings as shown in figure 1.

4.3 Hierarchical Transformer for Clinical Document Time Series

Instead of simply concatenating document embeddings within the same observation window to represent document time series, we additionally propose a hierarchical transformer to 1) account for the interactions between individual clinical notes via attention 2) achieve cross-modal time awareness by aligning clinical text embedding with correspond-

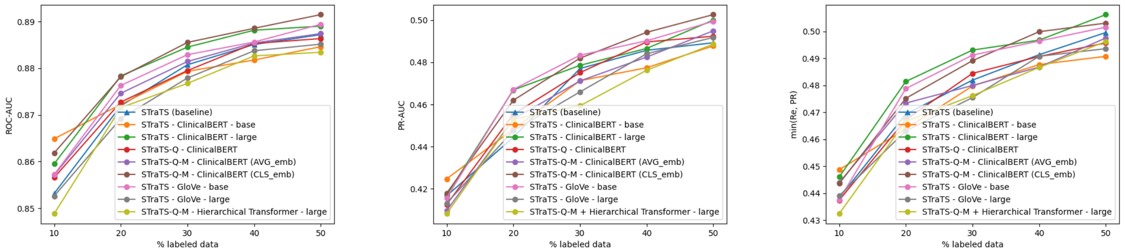

Figure 3: Sepsis prediction performance on MIMIC-III dataset for different percentages of labeled data averaged over 10 runs.

ing time embedding. As shown in figure 2, we encode time consistently with the learnable continuous value embedding module, and add it to corresponding document-level ClinicalBERT embedding for each clinical note in the observation windows, where the time encoding functions similarly to positional encoding in Vaswani et al. (2017); Dai et al. (2022). The time-aware text embeddings are then passed through transformer blocks and attention-based fusion module to claim interactions across individual clinical notes. By this practice, it brings cross-modal time-awareness to the entire multimodal learning framework via consistent time encoding.

5 Results

5.1 Clinical Time Series Forecasting

We train forecasting models with 2-hour forecasting window following each observation window on unsupervised data. We define observation windows with varied lengths: $\{min(0, x-24), x)|20 \leq x \leq 124, x\%4 = 0\}$. We evaluate the models with masked MSE (mean squared error), where the binary mask indicates the availability of ground truth in data. In addition to evaluating Multimodal STraTS-Q-M - ClinicalBERT against the baseline model, we conduct ablation studies to individually remove the quadruplet embedding module (revert back to monomodal triplet) and the text embedding module in late fusion concatenation. Furthermore, for experimental purposes, we also replace ClinicalBERT with generic GloVe model for text representation. Lastly, we replace the ClinicalBERT text embedding modules with our hierarchical transformer to represent clinical document time-series. We train base and large variations of the model, also further lower learning rate at pretraining to 0.0001 after 80 epochs from default learning rate (0.0005) due to the complexity of the model compared to others.

Table 3 shows the MMSE of the proposed models against baseline on test and validation data. It can be seen from the table that both the quadruplet embedding module and late fusion concatenation are able to individually improve model forecasting performance from baseline. With both combined, Multimodal STraTS-Q-M - ClinicalBERT reduces MMSE by 0.0981 from baseline on test data, achieving MMSE at as low as 5.1650. It is worth noting that when replacing CLS token embedding with average of all token embeddings as document-level representation, the same model underperforms baseline on test data and shows no noteworthy performance improvement at validation. In the meanwhile, the GloVe-based models (replacing ClinicalBERT with GloVe for text embedding in concatenation-based fusion model) are able to slightly outperform baseline at validation stage, whereas showing poor generalization to unseen test data and underperforms baseline by notable gap. Furthermore, by replacing the concatenation-based text embedding module with our hierarchical transformers, the large model with reduced learning rate is able to achieve the lowest MMSE on both test ($MMSE = 5.1535$) and validation data ($MMSE = 5.0038$), decreasing MMSE from baseline by 0.1096. This illustrates that our hierarchical transformer for clinical document time series is an effective approach compared to the simple concatenation of document-level embeddings.

We further run randomization tests on the outperforming models against baseline. As shown in table 4, we observe most of the p-values are below $\alpha - level$ ($p < \alpha$, $\alpha = 0.05$) with the exception to STraTS-Q-M - Hierarchical Transformer - large ($p = 0.69$), which is consistent with the marginal gap in MMSE of the model against baseline. The significance test results demonstrate that the majority of the outperforming models are significantly better than baseline in forecasting stage on test data.

5.2 24-h Sepsis Prediction with Labelled Data

As discussed in previous sections, our forecasting models can be used for early sepsis prediction in two ways: 1) directly fine-tuned on supervised data to predict sepsis 2) produce forecast on clinical variables to support rule-based implementations. In this work, we fine-tune the forecasting models with labelled sepsis patient data to illustrate the case of 24-hour sepsis prediction.

Figure 3 shows the ROC-AUC, PR-AUC and min(Re, Pr) (maximum of minimum of recall and precision across all thresholds). Multimodal STraTS-Q-M - ClinicalBERT is able to stably outperform baseline across different percentages of labelled data also on the downstream prediction task in a fully data-driven setup. While the hierarchical transformer model showed best performance on forecasting, it performs poorly after fine-tuning with labelled data on sepsis prediction. This observation is consistent with the arguements in Kaddour et al. (2022); Liu et al. (2023); Kaddour et al. (2023) that pretraining loss does not always correlate well with downstream performance.

6 Conclusion

In this work, we propose a multimodal transformer to incorporate both physiological time series and associated clinical notes from EHR data for clinical time series forecasting. We approach predictive tasks in the clinical domain primarily from a cause-prediction perspective, which allows our forecasting models to flexibly assist different clinical prediction tasks with rule-based checks in interpretable ways to practitioners in the field. We base our models on a strong monomodal baseline, and improved the model via meaningful multimodal fusion through integrating clinical text embedding modules. We additionally propose hierarchical transformers to represent clinical document time series using attention and time encoding. We conduct comprehensive experiments on MIMIC-III data primarily on forecasting, and observed that our multimodal models are able to significantly outperform baseline by notable gaps in MMSE. Our ablation studies illustrate that the atomic approaches in our multimodal fusion method (quadruplet embedding and late fusion via concatenation) are both able to individually improve model performance on forecasting, and achieve even more superior performance with both combined. Via integrating the hierarchical transformers, the forecasting model is able to further reduce MMSE with proper training setup, illustrating the effectiveness of our proposed hirarchical transformers for clinical document time series representation. Additionally, we fine-tune the forecasting models with supervised data for sepsis prediction, observing that most of the multimodal models are able to consistently outperform baseline on the downstream prediction task in a fully data-driven setup. While our models are based on encoder-only architectures, for future work we intend to explore multimodal encoder-decoder and decoder-only architectures with longer forecasting window. Meanwhile, we seek to reduce model parameters and enhance preprocessing steps in clinical note encoding procedures in future work.

7 Limitations

Despite the significant performance improvements over the baseline, our models generally have a higher number of parameters, resulting in increased computational costs. Additionally, our evaluation was conducted on a single dataset, assessing performance across multiple datasets would provide more robust and generalizable insights. Furthermore, our best-performing forecasting model did not consistently outperform the baseline during fine-tuning, indicating potential aspects for refinement.

8 Acknowledgements

We thank the reviewers for their comments.

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.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 65–74, New York, NY, USA. Association for Computing Machinery.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. 2020. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4353–4363. PMLR.

Alistair EW 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. *Scientific data*, 3(1):1–9.

Jean Kaddour, Oscar Key, Piotr Nawrot, Pasquale Minervini, and Matt J. Kusner. 2023. No Train No Gain: Revisiting Efficient Training Algorithms For Transformer-based Language Models. *arXiv preprint*. ArXiv:2307.06440 [cs].

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. 2022. When do flat minima optimizers work? In *Advances in Neural Information Processing Systems*, volume 35, pages 16577–16595. Curran Associates, Inc.

Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100:100057.

Anand Kumar, Daniel Roberts, Kenneth E Wood, Bruce Light, Joseph E Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, et al. 2006. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 34(6):1589–1596.

Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. 2023. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. *arXiv preprint arXiv:2305.02504*.

Steven Cheng-Xian Li and Benjamin M Marlin. 2016. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. *Advances in neural information processing systems*, 29.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR.

Vincent X Liu, Vikram Fielding-Singh, John D Greene, Jennifer M Baker, Theodore J Iwashyna, Jay Bhattacharya, and Gabriel J Escobar. 2017. The timing of early antibiotics and hospital mortality in sepsis. *American journal of respiratory and critical care medicine*, 196(7):856–863.

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.

Zhengdong Lu, Todd K Leen, Yonghong Huang, and Deniz Erdogmus. 2008. A reproducing kernel hilbert space framework for pairwise time series distances. In *Proceedings of the 25th international conference on Machine learning*, pages 624–631.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. 2022. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719. American Medical Informatics Association.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213. Curran Associates, Inc.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10.

Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. 2020. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical care medicine*, 48(2):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, et al. 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.

Christopher W Seymour, Foster Gesten, Hallie C Prescott, Marcus E Friedrich, Theodore J Iwashyna, Gary S Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M Terry, and Mitchell M Levy. 2017. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244.

Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. 2016. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):762–774.

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, et al. 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

Michael Staniek, Marius Fracarolli, Michael Hagmann, and Stefan Riezler. 2024. Early prediction of causes (not effects) in healthcare by long-term clinical time series forecasting. volume 252, pages 1–30.

Sindhu Tipirneni and Chandan K. Reddy. 2022. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data*, 16(6).

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. 2022. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *arXiv preprint arXiv:2203.14469*.

WHO. 2020. World health assembly 70, resolution 70.7: improving the prevention, diagnosis and clinical management of sepsis. 2017.

Jinghua Xu, Natalia Minakova, Pablo Ortega Sanchez, and Stefan Riezler. 2023. Early prediction of sepsis using time series forecasting. In *2023 IEEE 19th International Conference on e-Science (e-Science)*, pages 1–9.