TEMPORAL CROSS-ATTENTION FOR DYNAMIC EMBEDDING AND TOKENIZATION OF MULTIMODAL ELECTRONIC HEALTH RECORDS

Yingbo Ma, Suraj Kolla, Dhruv Kaliraman, Victoria Nolan, Zhenhong Hu,
Ziyuan Guan, Yuanfang Ren, Brooke Armfield, Tezcan Ozrazgat-Baslanti
Department of Medicine, University of Florida
{yingbo.ma,n.kolla,dhruv.kaliram,vnolan,hzhuf}@ufl.edu
{ziyuan.guan,renyuanfang,barmfield,tezcan}@ufl.edu

Tyler J. Loftus
Department of Surgery, University of Florida
tloftus@ufl.edu

Parisa Rashidi
Department of Biomedical Engineering, University of Florida
parisa.rashidi@ufl.edu

Azra Bihorac*, **Benjamin Shickel***
Department of Medicine, University of Florida
{abihorac,shickelb}@ufl.edu

ABSTRACT

The breadth, scale, and temporal granularity of modern electronic health records (EHR) systems offers great potential for estimating personalized and contextual patient health trajectories using sequential deep learning. However, learning useful representations of EHR data is challenging due to its high dimensionality, sparsity, multimodality, irregular and variable-specific recording frequency, and timestamp duplication when multiple measurements are recorded simultaneously. Although recent efforts to fuse structured EHR and unstructured clinical notes suggest the potential for more accurate prediction of clinical outcomes, less focus has been placed on EHR embedding approaches that directly address temporal EHR challenges by learning time-aware representations from multimodal patient time series. In this paper, we introduce a dynamic embedding and tokenization framework for precise representation of multimodal clinical time series that combines novel methods for encoding time and sequential position with temporal cross-attention. Our embedding and tokenization framework, when integrated into a multitask transformer classifier with sliding window attention, outperformed baseline approaches on the exemplar task of predicting the occurrence of nine postoperative complications of more than 120,000 major inpatient surgeries using multimodal data from three hospitals and two academic health centers in the United States.

1 INTRODUCTION

Electronic health records (EHRs) contain important information about patient encounters that support real-world healthcare delivery, and while artificial intelligence and machine learning have the theoretical potential to support clinical decision-making based on contextual representations of pa-

*Authors contributed equally

tient data, modeling real-world EHR time series is challenging due to its high dimensionality, multimodality, and temporal characteristics of the healthcare domain.

While evidence suggests that sequential deep learning approaches can outperform conventional machine learning for patient-level predictions (Shickel et al., 2023; Adiyeke et al., 2023; Morid et al., 2023), popular approaches such as recurrent neural networks (RNN) with long short-term memory (LSTM) (Memory, 2010) and gated recurrent networks (Chung et al., 2014) do not account for the temporal complexities of EHR data and may be suboptimal when learning temporal dynamics of patient health trajectories. Recently, transformers have been used for modeling temporal EHR data (Li et al., 2020; Shickel et al., 2022; Tipirneni & Reddy, 2022) and have been established as state-of-the-art approaches for predicting clinical outcomes from patient data sequences.

However, additional challenges persist when modeling EHR data with transformers, such as capturing temporal dependency across very long sequences (Li et al., 2022a) and modeling heterogeneous dependencies across variables (Zhang & Yan, 2022). Unstructured clinical notes, which contain important information about a patient encounter (Jensen et al., 2017), have the potential to provide added context to structured EHR. Recent studies have shown performance improvements obtained from jointly modeling structured EHR data and unstructured clinical notes for various multimodal clinical prediction tasks (Liu et al., 2022; Zhang et al., 2023). However, how to effectively learn the multimodal EHR representations in light of temporal EHR complexities remains an open question.

In this paper, we introduce a dynamic embedding and tokenization scheme to enable transformers to adapt to the unique challenges of multimodal clinical time series. Our scheme uses flexible positional encoding and a learnable time embedding to address the challenge of sparsity and irregular sampling, and a variable-specific encoding strategy for capturing distinct characteristics and relationships between temporal variables. To effectively combine structured, numerical time series data and free-text clinical notes, we adopted a cross-attention-based approach that learns a joint multimodal temporal representation. We demonstrate the effectiveness of our approach and analyze the relative contributions of each component of our framework using the benchmark task of predicting the onset of multiple postoperative complications following major inpatient surgery.

2 METHODS

Our dynamic embedding and tokenization framework for multimodal clinical time series includes methods designed to address (a) the temporal complexities of real-world EHR, and (b) the multimodal integration of structured EHR and unstructured clinical notes.

2.1 ENCODING DYNAMIC TEMPORALITY IN CLINICAL TIME SERIES

Figure 1 shows an overview of our dynamic embedding and tokenization scheme, which introduces three novelties to existing approaches: a flexible positional encoding, a learnable time encoding, and variable-specific encoding.

Flexible positional encoding. Multivariate EHR time series contain variables measured at different frequencies (e.g., a vital sign saved every minute vs. a laboratory test taken every 24 hours), and other variables that are measured at exactly the same time (e.g., blood pressure, heart rate, and respiratory rate from a bedside monitor). Traditional approaches that enforce a single resampled time interval, or present duplicate-time inputs in an arbitrary ordering, have the potential to lose information or inject unnecessary bias into the data representation. Our approach uses non-unique absolute positional indices based on the recorded timestamps so that variable tokens measured at the same time will be assigned the same positional index. To model the relationships between clusters of short-term activity across a long timeframe, we add a relative positional encoding to each token embedding (Shaw et al., 2018), which can help capture local token dependencies, especially for processing long sequences(Zaheer et al., 2020; Wei et al., 2021).

Learnable time encoding. Positional embeddings are used by models such as transformers to inject information about sequential order into time series representations (Dufter et al., 2022). However, positional embeddings alone omit critical information about the relative time between events. For applications of transformers to time series, time embeddings can help capture important temporal patterns (Zhang et al., 2020; Zeng et al., 2023). Our dynamic tokenization scheme uses Time2Vec

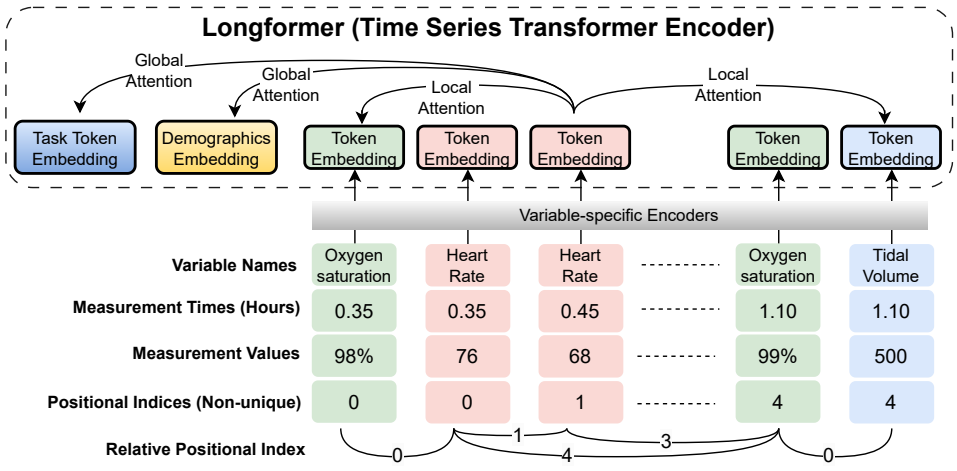

Figure 1: Dynamic embedding and tokenization scheme for multivariate clinical time series.

(Kazemi et al., 2019) to learn a model-agnostic vector representation for time. In Time2Vec, a time t is encoded to generate one non-periodic $\omega_{np}t+\phi_{np}$, and one periodic $sin(\omega_p t+\phi_p)$ time dependent vector, where ω and ϕ are learnable parameters (Liang et al., 2023).

Variable-specific encoding. A multivariate clinical time series often includes different categories of health variables (e.g., vital signs, laboratory tests, medications) that tend to exhibit distinct characteristics, numerical ranges, and temporal patterns. In recent EHR transformer implementations like BEHRT (Li et al., 2020)), all tokens within the sequence are embedded with the same encoder, which does not account for the heterogeneity among different variables. To address these existing challenges, we propose to use a separate encoder for each clinical variable for *intra-variable* temporal dynamics, and then concatenate the outputs of the seperate encoders to learn the *inter-variable* correlation and dependencies.

2.2 Cross-Attention for Joint Learning from Multimodal Clinical Time Series

To learn multimodal representations, we merged the embeddings of structured clinical time series and clinical notes using a validated cross-attention-based approach (Lu et al., 2019; Husmann et al., 2022; Liu et al., 2023; Zhang et al., 2023). Notes were encoded using a pretrained Longformer model(Li et al., 2022b), which outperformed other pretrained text models such as ClinicalBERT.

Consider the embeddings of clinical notes and structured EHR time series, denoted by X_{note} and X_{time}, respectively. The core of this approach lies in generating enriched feature sequence by searching relevant information between modalities. For example, $W_{time\rightarrow note} = softmax(\frac{Q_{note}K_{time}^T}{\sqrt{d_k}})$ represent a scoring matrix, whose (i,j)-th element measures the attention given by the information from the i-th time step from modality X_{note} and the j-th time step from modality X_{time}.

3 Experiments

In this section, we describe the evaluation of our approach on the benchmark task of predicting multiple in-hospital complications of major inpatient surgery using a real-world EHR dataset. Some technical details have been omitted for brevity and can be found in the Appendix.

3.1 Dataset

Our data consists of complete EHR records for all major inpatient surgeries occurring at three hospitals split among three hospitals (UF Health Gainseville, UF Health Jacksonville, and UF Health

North Jacksonville) between 2014 and 2019. The combined cohort consisted of 113,953 adult patients who underwent 124,777 inpatient surgeries.

For each inpatient surgery, our dataset consists of: (1) the patient's demographic and admission information, such as age, sex, and body mass index; (2) 14 intraoperative time series consisting of vital signs such as blood pressure, heart rate, and body temperature; (3) all preoperative and intraopertive clinical notes for an encounter, such as History and Physical (H&P notes) and operative reports; and (4) 9 binary labels indicating the occurrence of 9 postoperative complications.

3.2 Benchmark Multitask Classification

The goal is to predict the onset of nine postoperative complications following major inpatient surgery: prolonged ($>$ 48 hours) intensive care unit (ICU) stay, acute kidney injury (AKI), prolonged mechanical ventilation (MV), wound complications, neurological complications, sepsis, cardiovascular complications, venous thromboembolism (VTE), and in-hospital mortality. Models are trained on data available in the EHR up to the recorded timestamp of surgery end.

3.3 Models

Our embedding and tokenization scheme was designed for use with clinical transformers, and our primary classification model is a Longformer with sliding window attention. We compared the model trained with our dynamic embedding and tokenization scheme with several widely adopted baselines:

Tokenized gated recurrent units (GRUs) with attention: the tokenized sequential data was provided as input to a multi-layer GRU, followed by a self-attention layer (Tan et al., 2020; Shi et al., 2021).

Tokenized XGBoost: an XGBoost model was trained on tokenized sequential data (Wang et al., 2020a; Liu et al., 2022).

BHERT(Li et al., 2020), a widely used baseline transformer model for EHR data. In this baseline model, we removed variable-specific encoders and used one single embedding layer to encode sequential data. We also removed the relative positional embedding and time embedding, as they were not included in BHERT tokenization scheme.

Hi-BEHRT (Li et al., 2022a), a hierarchical transformer-based model for EHR data, which employs a sliding window to partition the long sequence into smaller segments. Within each segment, a transformer was utilized as a local feature extractor to capture temporal dynamics.

Self-supervised Transformer for Time-Series (STraTS) (Tipirneni & Reddy, 2022). STraTS uses a unique transformer to encode each variable with transformers and then uses a self-attention layer to generate the time-series embedding.

4 Results and Discussion

4.1 Time Series Modeling

Table 1: Model performance learning from clinical time series. Comparing the AUROC scores of different approaches for predicting nine postoperative outcomes.

Task	Mean	ICU	AKI	MV	Mortality	Wound	Neurological	Sepsis	Cardiovascular	VTE
GRU + Attention	0.771±0.04	0.857	0.718	0.783	0.816	0.712	0.753	0.791	0.762	0.747
XGBoost	0.765±0.03	0.851	0.716	0.771	0.815	0.709	0.748	0.788	0.760	0.727
Transformer (BEHRT)	0.749±0.01	0.843	0.701	0.765	0.800	0.701	0.725	0.770	0.748	0.699
Hi-BEHRT	0.781±0.03	0.863	0.730	0.789	0.835	0.721	0.769	0.801	0.780	0.769
STraTS	0.797±0.04	0.881	0.742	0.803	**0.857**	0.734	0.797	**0.813**	0.791	0.772
Transformer (Ours)	**0.801**±0.02	**0.883**	**0.749**	**0.810**	0.853	**0.739**	**0.800**	0.811	**0.797**	**0.774**

Table 1 compares the area under the receiver operating characteristic curve (AUROC) for our benchmark multitask classification task. As shown in the table, our dynamic tokenization scheme-based transformer model outperform all baseline models with the highest mean AUROC of 0.801. We

experimented with different types of variable-specific encoders (1-D convolutional layers (Kiranyaz et al., 2021) and transformer layers (Tipirneni & Reddy, 2022)) and found similar results (For 1-D CNN, transformer, and linear, the mean AUROC score was 0.796, 0.800, and 0.798, respectively). STraTS (Tipirneni & Reddy, 2022) slightly underperformed our approach, suggesting the utility of the added relative positional embeddings used by our approach. With the same tokenized sequence, GRU + Attention (mean AUROC: 0.771) outperformed transformer models with traditional tokenization scheme (mean AUROC: 0.749), indicating the advantages offered by our embedding and tokenization framework for other non-transformer modeling approaches.

4.2 Multimodal Fusion

Table 2: Model performance using multimodal fusion of structured time series and free-text clinical notes. Shown are the AUROC of different approaches for predicting nine postoperative outcomes.

Task	Mean	ICU	AKI	MV	Mortality	Wound	Neurological	Sepsis	Cardiovascular	VTE
Time Series Only	0.801±0.02	0.883	0.749	0.810	0.853	0.739	0.800	0.811	0.797	0.774
Clinical Notes Only	0.821±0.02	0.868	0.756	0.815	0.883	0.758	0.836	0.869	0.796	0.823
Late Weighted Fusion	0.813±0.03	0.882	0.748	0.807	0.850	0.752	0.815	0.809	0.797	0.778
Crossmodal Fusion + Concat2	0.822±0.01	0.866	0.755	0.816	0.881	0.754	0.831	0.867	0.797	0.831
Concat + Crossmodal Fusion3	**0.845**±0.03	**0.908**	**0.781**	**0.845**	**0.905**	**0.780**	**0.855**	**0.882**	**0.823**	**0.838**

Table 2 compares the AUROC of different models for multimodal learning of structured time series and free-text clinical notes. All of the multimodal models (time series + clinical notes) outperformed unimodal models utilizing either time series or clinical notes, aligning with conclusions from prior work(Husmann et al., 2022; Lyu et al., 2022). Concat + Crossmodal Fusion (shown in Appendix, Figure 2) performed the best, establishing a state-of-the-art mean AUROC of 0.845 for this task. Multimodal models trained without cross-attention resulted in less accurate predictions, suggesting this cross-attention-based fusion approach can effectively learn joint multimodal representations of time series data and clinical notes.

5 Conclusion

In this work, we introduced a dynamic embedding and tokenization scheme to adapt to the unique temporal challenges found in multimodal clinical time series. Experiments with real-world EHR databases on a benchmark clinical prediction task highlighted its advantages. Our work makes several contributions to the ongoing research of clinical time series modeling, as well as exploring innovative approaches to incorporate diverse health data sources.

A Appendix

Below is the appendex section, including multimodal architectures, dataset statistics, data preprocessing details, experiment details, and related work.

A.1 Multimodal Architectures

Please see Figure 2 and Figure 3 for two cross-attention-based fusion architectures.

A.2 Dataset Statistics

For each patient who underwent surgeries, we extracted following features:

1. 9 preoperative demographic and admission information from 113,953 patients, including age (Mean 51 y, Min 18 y, Max 106 y), sex (48% male, 52% female), language, ethnicity, race, smoking status, zip code, and body mass index.
2. 14 intraoperative temporal vital signs, including systolic blood pressure, diastolic blood pressure, mean arterial pressure, heart rate, respiratory rate, oxygen flow rate, fraction of inspired oxygen (FIO2), oxygen saturation (SPO2), end-tidal carbon dioxide (ETCO2),

minimum alveolar concentration (MAC), positive end-expiratory pressure (PEEP), peak inspiratory pressure (PIP), tidal volume, and body temperature.

3. 173 types of all preoperative and intraopertive clinical notes for an encounter, such as History and Physical (H&P notes) and operative reports.
4. 9 clinical outcomes, the incidence of complications include 23.29% ICU stay (for 48 h or more), 13.09% acute kidney injury, 8.64% prolonged mechanical ventilation, 2.00% in-hospital mortality, 13.48% wound complications, 15.09% neurological complications, 8.20% sepsis, 12.18% cardiovascular complications, and 4.51% venous thromboembolism.

A.3 Data Preprocessing

Demographic and Admission Information For categorical data, we converted each to one-hot vectors, and concatenated with remaining numerical values. Missing static features was imputed with cohort medians.

Time Series Data For 14 intraoperative time series data, their variable names were converted to unique integer identifiers; the measured values for each variable was were normalized to zero mean and unit variance based on the values from the training set; their measurement time, in the format of "month/day/year hour:min:sec", were first converted to unix timestamps and then also normalized similarly. For absolute positional indices, we assign one integer positional index for each token yet not enforcing the restriction that positional indices are unique and if different variables were measured at the same time. For relative positional embeddings, we generated the relative positional representation based on the GitHub code[1] for the original paper (Shaw et al., 2018).

Clinical Notes In the preprocessing phase, we merged all types of notes per surgery, converted the text to lowercase, and removed special characters and de-identification placeholders. Subsequently, we generated embeddings by first tokenizing the whole text using the clinically pretrained tokenizer. The tokens were then chunked to fit the pretrained clinical LLM, and the last hidden layer output for the CLS token was extracted as the embedding for each chunk. The final representation for each surgery was obtained by calculating the average of all these embeddings. We fixated on the Clinical Longformer [2] for generating the embeddings due to its superior performance in classifying with clinical notes, following extensive testing with various models from Huggingface including BioBERT [3], BiomedBERT[4], ClinicalBERT [5], and Clinical Longformer[6].

[1] https://github.com/microsoft/MPNet/blob/master/pretraining/fairseq/modules/rel_multihead_attention.py
[2] https://huggingface.co/yikuan8/Clinical-Longformer
[3] https://huggingface.co/dmis-lab/biobert-base-cased-v1.2
[4] https://huggingface.co/allenai/biomed_roberta_base
[5] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
[6] https://huggingface.co/yikuan8/Clinical-Longformer

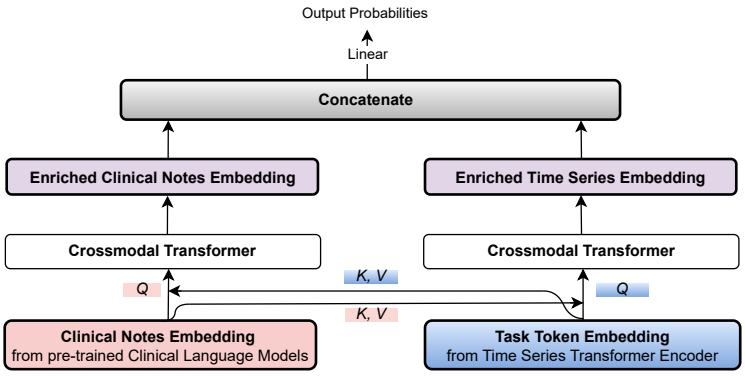

Figure 2: Overview of Crossmodal Fusion + Concat.

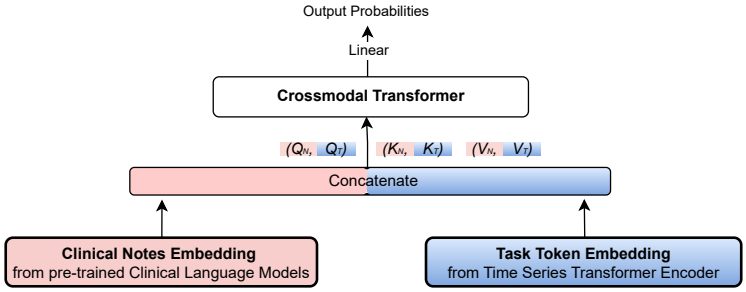

Figure 3: Overview of Concat + Crossmodal Fusion.

A.4 Experiment Details

Multi-task Training Our model was trained with the multi-task fashion for predicting 9 postoperative outcomes. To do this, we expanded the notion of "[CLS]" token for text classification and prepended 9 global tokens to our tokenized sequences, one for each of our postoperative outcomes, so that self-attentions were computed among all sequence elements for each clinical outcome token.

Longformer for processing long sequences Vanilla transformer models are largely limited by their capacity in processing long sequences (*max_sequence* = 512) because of its quadratic complexity with respect to the sequence length. Therefore, in this paper we used Longformer (Beltagy et al., 2020) for time series modeling transformers. Longformer architecture introduces sliding window attention that allows the model to attend to only a subset of tokens within a window, reducing the overall computational complexity. This design makes Longformer well-suited for processing longer sequences (*max_sequence* = 4096).

Experimental Setup We trained the models on two NVIDIA A100-SXM4-80GB GPUs for 30 epochs to leverage hardware acceleration. We used a batch size of 32 per GPU for the best performing model.

Hyperparameters We used the following hyperparameters for optimization and regularization, Adam optimizer with a learning rate of 1e-4, dropout of 0.2, and weight decay of 1e-4. For the transformer models, including the Longformer, we limited the models to only 1 attention head and 1 layer per head, as this configuration produced the best results.

Loss Function We chose to use a Binary Cross-Entropy with Logits (BCEWithLogitsLoss) with the parameter *pos_weight* set to the weight of the positive class for each prediction outcome, given the unbalanced nature of our dataset. This allows models to be more sensitive to minority class by increasing the cost of misclassification of it.

A.5 Related Work

Transformers for EHRs Recent transformers for EHRs focused on adapting vanilla transformer architecture for tokenizing irregular sampling and learning temporal dependencies in long sequences. For example, for sparsity and irregular sampling, Zhang et al. (2021) used graph neural network to embed irregularly sampled and multivariate time series, capturing sensor dynamics from observational data. Tipirneni & Reddy (2022) treated time-series as sets of observation triplets and uses transformers to encode continuous time and variable values, eliminating the need for discretization. For processing long sequences, Li et al. (2022a) employed a sliding window to partition the complete medical history into smaller segments. Within each segment, a Transformer was utilized as a local feature extractor to capture temporal interactions. Motivated by this line of research, we introduced our dynamic embedding approache in this paper, aiming to address temporal EHR challenges by learning time-aware representations patient time series.

Multimodal Representation Learning for Health Combining diverse sources of data sources in medical domain is promising for more comprehensive understanding of patients' health conditions (Raghupathi & Raghupathi, 2014), more accurate health outcome predictions (Shickel & Bihorac, 2023), and building next-generation foundational medical models for generative AI (Moor et al.,

2023). The core of this research effort is multimodal representation learning where all the modalities are projected to a common space while preserving information from the given modalities (Liang et al., 2022). Traditional data fusion methods, such as early and late fusion, are insufficient to learn the correlations and dependencies among different modalities (Ma et al., 2022). Recently, transformer-based architecture, thanks to its superior ability to capture cross-modal interactions by self-attention and its variants (Xu et al., 2023), has achieved great success in various multimodal machine learning tasks in different domains, such as multimodal action recognition (Wang et al., 2020b), image segmentation (Xiao et al., 2023), and affect detection (Ma et al., 2023).

ACKNOWLEDGMENTS

We would like to thank the NVIDIA Corporation for providing computational resources used for this research. This research was also supported by National Institute of Health (NIH) through grant R01 GM110240. Any opinions, findings, conclusions, or recommendations expressed in this research are those of the authors, and do not necessarily represent the official views, opinions, or policy of NIH.

REFERENCES

Esra Adiyeke, Yuanfang Ren, Matthew M Ruppert, Benjamin Shickel, Sandra L Kane-Gill, Raghavan Murugan, Parisa Rashidi, Azra Bihorac, and Tezcan Ozrazgat-Baslanti. A deep learning–based dynamic model for predicting acute kidney injury risk severity in postoperative patients. *Surgery*, 2023.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763, 2022.

Severin Husmann, Hugo Yèche, Gunnar Rätsch, and Rita Kuznetsova. On the importance of clinical notes in multi-modal learning for ehr data. *arXiv preprint arXiv:2212.03044*, 2022.

Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific reports*, 7(1):46226, 2017.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022a.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022b.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1559–1568, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*, 2022.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics*, 145:104466, 2023.

Shuhui Liu, Bo Fu, Wen Wang, Mei Liu, and Xin Sun. Dynamic sepsis prediction for intensive care unit patients using xgboost-based model with novel time-dependent features. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4258–4269, 2022.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

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

Yingbo Ma, Mehmet Celepkolu, and Kristy Elizabeth Boyer. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pp. 45–55, 2022.

Yingbo Ma, Mehmet Celepkolu, Kristy Elizabeth Boyer, Collin F Lynch, Eric Wiebe, and Maya Israel. How noisy is too noisy? the impact of data noise on multimodal recognition of confusion and conflict during collaborative learning. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 326–335, 2023.

Long Short-Term Memory. Long short-term memory. *Neural computation*, 9(8):1735–1780, 2010.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Mohammad Amin Morid, Olivia R Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems*, 14(1):1–29, 2023.

Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2:1–10, 2014.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Zhenkun Shi, Sen Wang, Lin Yue, Lixin Pang, Xianglin Zuo, Wanli Zuo, and Xue Li. Deep dynamic imputation of clinical time series for mortality prediction. *Information Sciences*, 579:607–622, 2021.

Benjamin Shickel and Azra Bihorac. The dawn of multimodal artificial intelligence in nephrology. *Nature Reviews Nephrology*, pp. 1–2, 2023.

Benjamin Shickel, Brandon Silva, Tezcan Ozrazgat-Baslanti, Yuanfang Ren, Kia Khezeli, Ziyuan Guan, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. Multi-dimensional patient acuity estimation with longitudinal ehr tokenization and flexible transformer networks. *Frontiers in Digital Health*, 4:1029191, 2022.

Benjamin Shickel, Tyler J Loftus, Matthew Ruppert, Gilbert R Upchurch Jr, Tezcan Ozrazgat-Baslanti, Parisa Rashidi, and Azra Bihorac. Dynamic predictions of postoperative complications from explainable, uncertainty-aware, and multi-task deep neural networks. *Scientific Reports*, 13 (1):1224, 2023.

Qingxiong Tan, Mang Ye, Baoyao Yang, Siqi Liu, Andy Jinhua Ma, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and PongChi Yuen. Data-gru: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 930–937, 2020.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Yuan Wang, Yake Wei, Hao Yang, Jingwei Li, Yubo Zhou, and Qin Wu. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. *BMC Medical Informatics and Decision Making*, 20(1):1–13, 2020a.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, pp. 2514–2520, 2020b.

Wei Wei, Zanbo Wang, Xianling Mao, Guangyou Zhou, Pan Zhou, and Sheng Jiang. Position-aware self-attention based neural sequence labeling. *Pattern Recognition*, 110:107636, 2021.

Hanguang Xiao, Li Li, Qiyuan Liu, Xiuhong Zhu, and Qihang Zhang. Transformers in medical image segmentation: A review. *Biomedical Signal Processing and Control*, 84:104791, 2023.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Machine learning for healthcare conference*, pp. 566–588. PMLR, 2020.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021.

Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pp. 41300–41313. PMLR, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.