2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995044

UniMed: Multimodal Multitask Learning for Medical Predictions

Xiongjun Zhao[1], Xiang Wang[2] Fenglei Yu[2] Jiandong Shang[3] Shaoliang Peng[12*]
[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]The Second Xiangya Hospital, Central South University, Changsha, China
[3]HeNan Supercomputer Center, ZhengZhou University, ZhengZhou, China
{xiongjunzhao, slpeng}@hnu.edu.cn, wangxiang@csu.edu.cn, shangjiandong@zzu.edu.cn

***Abstract*—Recently, deep learning techniques based on electronic health record (EHR) data have achieved success in medical prediction. However, due to the complexity, heterogeneity nature of EHR data, most previous studies build models based on single-modal data (e.g. the structured data or the unstructured free-text data). Although some studies have trained the models based on multimodal EHR data and achieved more advanced performance, they still suffer from the clinical practicability problems, as they require separate modeling for each medical prediction task. Moreover, they ignore the potential correlation between clinical prediction tasks. In this work, we propose UniMed, a *Uni*fied model handles multiple *Med*ical prediction tasks simultaneously by learning from multimodal EHR data. Our UniMed model encodes each input modality separately and uses a transformer decoder followed by task-specific prediction heads to predict each medical task. Experimental results conducted on publicly available EHR dataset demonstrate that there is a time-progressive correlation between medical prediction tasks and show the effectiveness of our method.**

***Index Terms*—Medical Predictions, Electronic Health Records, Multimodal, Multitask Learning**

I. Introduction

In recent years, deep learning techniques have achieved great performance in the medical domain, such as chest X-ray pneumonia diagnosis [1], skin cancer classification [2]. At the same time, with a large amount of patient-level EHR data being generated in hospitals every day, researchers began to focus on using EHR data to predict patient clinical outcomes, such as intensive care unit (ICU) mortality prediction [3], length-of-stay [4], and acute respiratory failure (ARF) prediction [5].

Unlike medical images and medical plain text data, real-world EHR data is longitudinal and multimodal, including structured data and unstructured data. For instance, Structured data in EHR can usually be divided into two modalities: time-invariant and time-dependent. Time-invariant data refers to static or discrete categories during hospitalization, such as patient gender, age and medication codes. Time-dependent data refers to dynamic or continuous features, such as vital signs and laboratory tests. Unstructured data refers to free-text, such as clinical notes. Therefore, the EHR data of patients can be summarized into the above three modalities.

Due to the complexity of EHR data, researchers only focused on structured data or unstructured free text data in previous work. Both [6] and [7] use time-dependent data to train a recurrent neural networks (RNN) for medical prediction. While [8] use transformer to model clinical notes for predicting hospital readmission.In general, most literature uses single-modality data as input to predict a single task. Recently, some research works have begun to enhance model prediction performance by fusing multimodal EHR data [9], [10]. Similarly, [11] propose multimodal fusion architecture search strategy for better leverage multimodal EHR data.

Despite the above achievements in medical prediction using deep learning for specific tasks, there has not been much effort to explore the potential correlation between medical tasks to improve prediction performance with multimodal EHR data. But in a clinical research, [12] shows that acute respiratory failure is related to a mortality rate of 35%-46% in clinic. Additionally, applying previous models to real clinical practice is a challenge, because it is necessary to build and train a model for each medical prediction task, even if their training data come from the same EHR system. Overall, as a step towards general intelligence, is it possible to use multimodal EHR data to build a unified single model that handles multiple tasks simultaneously, and take advantage of the potential correlation of medical tasks.

Inspired by the fact that physicians consider multiple data and previous clinical events when making decisions, we define the correlation between medical prediction tasks as a time-progressive, which means that the probability of the current task output can refer to the output of all previous tasks. As a step forward, we use multimodal multitask learning and build a simple but effective medical prediction model, UniMed. It takes three modalities of structured data and unstructured data as inputs and jointly train on multiple tasks. UniMed consists of three encoders that encode each input modality as a feature vector, and a transformer decoder over the encoded input modalities, and then applies task-specific prediction heads to the hidden states of the decoder to make the final prediction for each medical task. The self-attention mechanism of transformer can learn to focus on a main modality of the input for different tasks and consider previous outputs in current step. Compared to previous work on single-task learning, we train UniMed and achieve comparable performance to well-established prior work on variety of medical prediction tasks. Further analysis of the publicly available EHR dataset

[*] Corresponding author

demonstrates that applying multimodal multitask learning is a plausible way for medical predictions.

Our contributions in this paper can be summarized as follows:

- We first propose to apply multimodal multitask learning and explore the potential correlation of medical tasks for better medical predictions.
- We propose **UniMed**, a **uni**fied single model for **med**ical predictions that handles multiple input modalities and tasks simultaneously. Promoting general intelligence in the medical.
- Empirical evidence demonstrating that UniMed is superior to previous single-task models. By analyzing various time-related medical tasks, we further show that exploiting the time-progressive correlation between tasks and multitask learning can significantly improve prediction performance.

II. Related Works

Early attempts of medical predictions based on EHR data were single-task model [5], [10], [11]. Recently, some clinical studies have observed a positive correlation between medical events (tasks). [13] retrospectively analyzed the positive correlation between ARF and all-cause mortality based on the MIMIC-III [14] database. In addition, ARF is generally caused by respiratory diseases or pulmonary vascular diseases [15], which can assist in the subsequent diagnosis of diseases. Therefore, in this work, we use multitask learning and transformer based on multimodal EHR data to explore the correlation between tasks for better medical predictions.

For multitask learning, recent studies have shown that multitask learning achieves state-of-the-art performance in vision [16], language tasks [17], as well as in multimodal domain [18]. In the medical domain, several studies have been conducted to identify different subgroups of patients and use multitasking learning to share knowledge with these subgroups [19]. Other studies have explored the utility of using multitask learning to perform clinical prediction tasks with time series data [20]. The closest work [21] uses the transformer architecture to predict seven clinical tasks. However, none of these works consider problem settings where tasks may be correlated in sequential or temporal structure.

For Transformer in healthcare, there has been a lot of work training transformer models using clinical text [22], continuous variables [23], or discrete medical codes [24]. Following [25], we also train our UniMed as auto-regressive (AR), in which our model can learn all the target tasks in parallel.

III. Our Method

The overall architecture of our method, UniMed, is shown in Figure 1. It is built upon the encoder-decoder framework and contains four main components: a data preprocess pipeline to extract each modality data from raw EHR, encoders for each input modality type, a share transformer decoder, and task-specific prediction heads.

A. Data Preprocess

Our model considers three input modalities of structured and unstructured data: time-invariant data, time-dependent data and free-text. For structured data, we apply a flexible data-driven preprocessing pipeline (FIDDLE) [5] to extract time-invariant data and time-dependent data. Taking the raw EHR of an ICU stay as input, the time-invariant data can be represented as $\mathbf{x}_{ti} \in \mathcal{R}^{d_1}$, where d_1 is the dimension of the time-invariant data. Similarly, the time-dependent data can be represented as $\mathbf{X}_{td} \in \mathcal{R}^{t \times d_2}$ after FIDDLE, where t denotes the number of hours in ICU stay and d_2 is the dimension of time-dependent data. For unstructured data, which refer to free-text clinical notes in this work. We apply the wordpiece tokenization proposed by Google NMT [26] to split the original words into smaller-grained wordpieces. It can be represented as $\mathbf{x}_{cn} \in \mathcal{R}^{d_3}$, where d_3 is the length of the notes.

B. Encoders

Our encoding process is inspired by [10]. We encode them separately according to the nature of each input modality.

Time-invariant: The time-invariant data is discrete and static, such as demographic information. In our model, we use a fully connected network followed by the Relu activation function to encode the input $\mathbf{x}_{ti}$:

$$\mathbf{h}_{ti} = \text{Linear}(\mathbf{x}_{ti}) \tag{1}$$

$\mathbf{h}_{ti} \in \mathcal{R}^{d'_1}$, where $\mathbf{h}_{ti}$ is the encoded feature, which dimension is d'_1.

Time-dependent: Considering that the time-dependent data contains continuous data from clinical monitor and hourly features from lab tests. As a result, Long Short Term Memoey (LSTM) [27] with the ability to handle temporal sequences are applied to encode the input $\mathbf{x}_{td}$:

$$\mathbf{h}_{td} = \text{LSTM}(\mathbf{X}_{td}) \tag{2}$$

$\mathbf{h}_{td} \in \mathcal{R}^{d'_2}$, where $\mathbf{h}_{td}$ is the encoded feature, which dimension is d'_2.

Free-text Clinical Notes: In this work, we encode the free-text clinical notes input using Clinical-Bert [22], which pre-trained on large EHR corpora. Given the input notes, we use wordpiece tokenize them into a sequence tokens $\mathbf{x}_{cn} = \{x_1, x_2, \cdots, x_s\}$, with $x_1 = [CLS]$. Then the token sequence is used as input to ClinicalBert model to extract a sequence of textual feature.

$$\mathbf{h}_{cn} = \mathbf{h}_{[CLS]} = \text{ClinicalBert}(\mathbf{x}_{cn}) \tag{3}$$

$\mathbf{h}_{cn} \in \mathcal{R}_{d'_3}$, where $\mathbf{h}_{cn}$ is the encoded feature, which dimension is d'_3.

After encoding the input modalities, We align the dimensions of $\mathbf{h}_{ti}$, $\mathbf{h}_{td}$ and $\mathbf{h}_{cn}$ for subsequent concatenate them.

$$\mathbf{h}_{enc} = \text{concat}(L_{d'_1 \to e}(\mathbf{h}_{ti}), L_{d'_2 \to e}(\mathbf{h}_{td}), L_{d'_3 \to e}(\mathbf{h}_{cn})) \tag{4}$$

where L is a linear projection for dimension alignment. In our implementation, the dimension e of $\mathbf{h}_{enc}$ after alignment is set to 512.

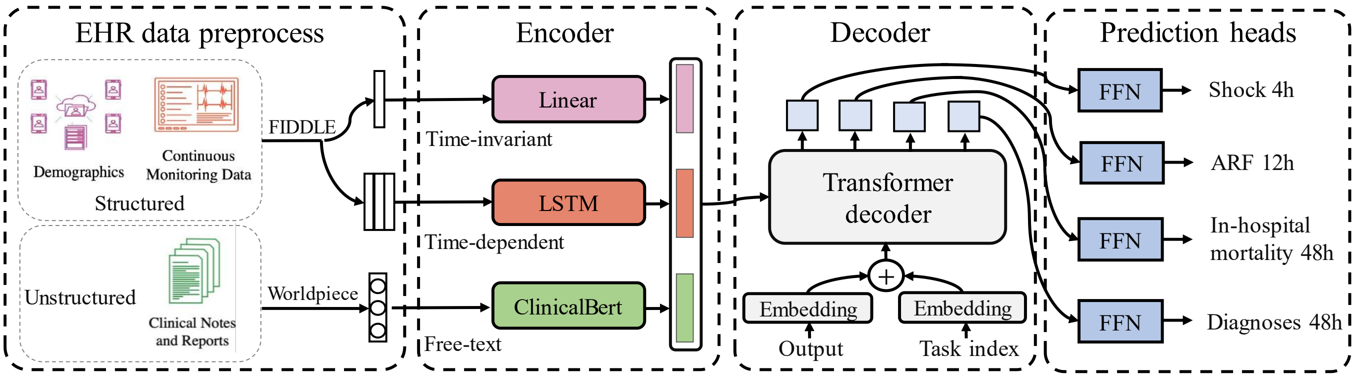

Fig. 1. The overview architecture of our proposed UniMed. UniMed takes multimodal EHR data as input and jointly handles multiple medical prediction tasks with a unified decoder-encoder framework.

C. Decoder

Inspired by sequential output and attention mechanisms in machine translation [28], the multitask medical prediction tasks can naturally be seen as an auto-regressive decoder task. Therefore, i n this work, our decoder D follows the standard architecture of the transformer [25] decoder with number of layers N_d.

Similar to the dictionary of language models, we build a dictionary with all possible label results in the prediction task. Since the transformer decoder is permutation-invariant, we need add some relative or absolute position information to maintain the order relationship between tasks. The positional encodings that we refer to as *task index*. In our implementation, we use two learned embeddings to convert the output and task-index tokens to vectors of length N and dimension d_{dec}, then sum up them to obtain the final embedding $\mathbf{q}$. The decoder D takes the embedding sequence $\mathbf{q}$ and the encoded input sequence $\mathbf{h}_{enc}$ as inputs and outputs a sequence of decoded hidden states $\mathbf{h}_{enc}$, which has the same length N as the embedding $\mathbf{q}$.

$$\mathbf{h}_{dec} = D(\mathbf{h}_{enc}, \mathbf{q}) \tag{5}$$

In the decoder, the masked attention is applied to the $\mathbf{q}$ and cross-attention is applied among the encoded input modalities $\mathbf{h}_{enc}$, which ensures the prediction task focus on the special information from the encoded input and the temporal relation between tasks.

D. Task-specific Prediction Heads

For each task, we use a special prediction head follow the decoder hidden states $\mathbf{h}_{dec}$. The prediction head is computed by a fully connected feed-forward network (FFN), which consists a 2-layer linear transformations with a ReLU activation. In our experiments, we only consider the classification tasks in medical prediction. For the binary classification task, we follow a linear layer with $sigmoid$ function after FFN:

$$\mathbf{p} = \text{sigmoid}(\text{FFN}(\mathbf{h}_{dec})) \tag{6}$$

where $\mathbf{p} \in \{0, 1\}$. For the multi-label tasks, we use $softmax$ function instead of $sigmoid$:

$$\mathbf{P} = \text{softmax}(\text{FFN}(\mathbf{h}_{dec})) \tag{7}$$

where $\mathbf{P} \in \mathcal{R}^M$ and M is the number of classes.

Following [5] and [10], we use the cross-entropy between the prediction and the ground-truth to compute the classification loss for all prediction tasks. Then, we simply add the loss of all tasks to get the overall loss:

$$\mathcal{L} = \sum_i \mathcal{L}_i \tag{8}$$

IV. Experiment Setup

A. Dataset

We conduct experiments on Medical Information Mart for Intensive Care (MIMIC-III) [14]. It is a large, single-center database that contains real-world EHR data for 53,423 hospital ICU stays, admitted to ICU from 2001 to 2012 at the Beth Israel Deaconess Medical Center. In our experiments, we focus on the data extracted from MetaVision system from 2008 and 2012, which contains 17,710 patients with 23,620 stays. Considering task labels and modality requirements, we filtered out patients younger than 18, lack of labels and those with incorrect or missing records. After data pre-processing, there are still 8,577 ICU stays. We represent medical histories of patient in time-invariant, time-dependent and free-text clinical notes, as described in the previous section. The data is randomly split into train, validation and test sets in an $8 : 1 : 1$ ratio.

B. Predict Tasks

For each ICU stay, we sought to sequentially predict shock at 4h, acute respiratory failure (ARF) at 12h, in-hospital mortality and diagnoses both at 48h.

- **Shock 4h:** is defined as inadequate perfusion of blood oxygen to organs or tissues [29] within 4 hours during this ICU stay. This is a binary classification problem, as the model needs to predict whether the patient will develop into shock at 4h.

- **ARF 12h:** is defined as the need for respiratory support with positive-pressure mechanical ventilation [29] within 12 hours during this ICU stay. This is a binary classification problem, as the model needs to predict whether the patient will develop into ARF at 12h.
- **In-hospital mortality 48h:** is defined based on patient outcome within 48 hours during this ICU stay. This is a binary classification problem, as the model needs to predict whether the patient will develop into death at 48h.
- **Diagnoses 48h:** is defined as the related diseases diagnosed by the doctor within 48 hours during this ICU stay. Each ICU stay may relate to multiple diseases, Following [10], we extracted 1024 disease groups from the MIMIC-III, and the diseases were identified by International Classification of Diseases, 9th Revision (ICD-9) [30], Because each ICU stay may be associated with multiple diseases, this is a multi-label classification task.

For binary classification problems, such as shock 4h, we use AUROC (Area Under the Receiver Operating Characteristic curve) as evaluation metric. For multi-label classification task, like diagnoses 48h, we use Top-k recall as evaluation metric, where $k = 10$.

C. Experimental Details

In our implementation, we use the Adam optimizer with a learning rate of $1e-4$. The dropout is set to 0.1. Following the data preprocessing and encoding, the dimensions d'_1, d'_2 and d'_3 of the encoded data are set to 64, 128 and 768 respectively. For transformer decoder, we use a small transformer network with $N_d = 3$ identical layers, 512 hidden size and 6 parallel attention heads.

D. Baselines

First, we select FIDDLE [5] as the baseline of single-modal and single-task. FIDDLE applies deep learning methods such as CNN and LSTM to predict. Then, we also select three competitive multimodal and single-task models proposed more recently, including LstmBert and BertEncoder proposed in [10], and MUFASA [11]. For multitask model, [21] uses structured EHR data for multitask prediction of clinical outcomes based on transformer. Because the code is not open source, we exclude it from our baselines. Besides, we add a single-task UniMed trained for each predictive task and a multi-task UniMed trained jointly for all tasks.

V. Results and Discussion

A. Main Results

Our main results are shown in Table I. First, In our experiments, the three models trained with multimodal EHR data outperformed FIDDLE, which uses only structured EHR data, on most tasks. Secondly, we select two best models in [10], LstmBert and BertEncoder, which take time-dependent as the main modality and clinical notes as the main modality, respectively. It can be seen from the experimental results that we reproduce the observation in [10]: for the shock, ARF and in-hospital modality tasks, the LstmBert model performs better, while the BertEncoder performs better in the diagnoses task. Third, for MUFASA, we follow the implementation in [11], which adopts the architecture of self-attention and different fusion strategies for each modality and achieves good performance in our experiments. For our UniMed, we experiment with single-task and multi-task, where single-task builds the network separately for each task, and multi-task jointly handles multiple tasks. From the experimental results, it can be concluded that multi-task perform better than single-task. In particular, there is a relatively large performance improvement for tasks with backward timestamps, such as a 5.4% increase in diagnoses 48h tasks.

B. Ablation Study

In this section, we discuss the effects of different input modalities and decoders on performance as shown in Table II and Table III. Then we further analyze the time-progressive correlation between different timestamps in the same task as shown in Figure 2.

To answer how much the performance of the model is affected by the input data, we train three models with the same network structure but with different inputs. First, we use only the structured data, including time-variant and time-dependent data for prediction. secondly, we use only the unstructured data including free-text clinical notes for prediction. From the Table II, we observe that the structure data is more effective than unstructured data in shock and ARF tasks. While the unstructured data is more important for diagnoses task. Benefiting from the cross-attention mechanism in transformer, UniMed adjusts the weights of input modalities according to the task and achieves better performance.

We further use two decoders, GRU and GRU with attention, to understand the effect of transformer decoder and multitask in our method. The GRU and GRU attention follow the implementation in Pytorch tutorials [31]. From the III, we observe that the three multi-task learning models are generally better than single-task learning, and the decoders with attention mechanism are obviously better. Furthermore, compared to these two GRU-based decoders, the transformer architecture appears to be more suitable for multimodal EHR data and multitasking, which has also been verified in other domains [18].

To further explore the correlation between tasks, we test the same task at different timestamps. We select ARF task at 4h and 12h, as well as shock at 4h and 12h. By observing our experimental results in Figure 2 and FIDDLE experiments [5], it can be concluded that the performance of single task model in ARF 12h and shock 12h tasks decreases because the time span of the two tasks is relatively long. Our UniMed jointly learns ARF 4h and ARF 12h, which can effectively improve the performance of ARF 12h task, as well as shock 12h.

C. Discussion

Sufficient experiments show that using multimodal multitask learning for medical predictions on EHR data is a potential and effective direction. More importantly, building

TABLE I
PERFORMANCE OF VARIOUS MODELS FOR MEDICAL PREDICTIONS. OUR UNIMED BASED ON MULTI-TASK TRAINING IS SUPERIOR TO THE SINGLE-TASK MODEL TRAINED ALONE IN MOST TASKS.

Model	Task	Shock 4h	ARF 12h	In-hospital mortality 48h	Diagnoses 48h
	Metric	AUROC	AUROC	AUROC	Recall@10
FIDDLE [5]	CNN	0.824	0.757	0.885	0.309
	LSTM	0.811	0.760	0.863	0.310
Bo Yang et al. [10]	LstmBert	0.843	0.779	0.867	0.303
	BertEncoder	0.811	0.717	0.860	0.331
MUFASA [11]		**0.848**	0.775	0.871	0.344
UniMed (ours)	Single-task	0.831	0.764	0.873	0.331
	Multi-task	0.836	**0.781**	**0.892**	**0.385**

TABLE II
PERFORMANCE OF INDEPENDENT MODALITY MODELING.

Modality	ARF 4h	ARF 12h	In-hospital mortality 48h	Diagnoses 48h
	AUROC	AUROC	AUROC	Recall@10
Structured data only	0.802	0.756	0.857	0.281
Unstructured data only	0.742	0.715	0.839	0.336
Both (UniMed)	**0.836**	**0.781**	**0.892**	**0.385**

TABLE III
PERFORMANCE OF DIFFERENT DECODERS IN SINGLE-TASK AND MULTI-TASK.

Decoder	Task	Shock 4h	ARF 12h	In-hospital mortality 48h	Diagnoses 48h
	Metric	AUROC	AUROC	AUROC	Recall@10
GRU	Single-task	0.815	0.746	0.865	0.316
	Multi-task	0.821	0.760	0.871	0.331
GRU with attention	Single-task	0.823	0.758	0.867	0.329
	Multi-task	0.836	0.770	0.886	0.363
Transformer (UniMed)	Single-task	0.831	0.764	0.873	0.331
	Multi-task	**0.836**	**0.781**	**0.892**	**0.385**

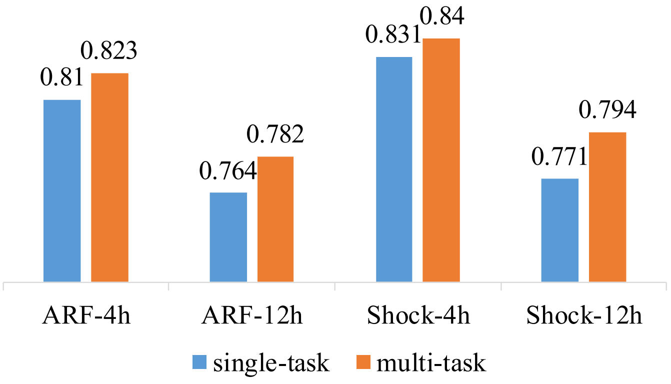

Fig. 2. Performance of different timestamps on the same medical prediction task in single-task and multi-task.

a unified predictive model for multiple clinical outcomes of patients can accelerate the application of artificial intelligence to real clinical practice. However, our method also has some limitations. First, our method needs to predict multiple tasks across different timestamps. Compared with the single-task model, our UniMed has a much smaller dimension of time-dependent data to avoid leakage of label data. For example, for the in-hospital mortality 48h task, the time-dependent data have more than 7000 dimensions in FIDDLE [5], but only about 4000 dimensions in UniMed, because the combined prediction of shock 4h and in-hospital mortality at 48h required exclusion of data between 4h and 48h. For this problem, incremental learning could be used to solve it, which is a significant and meaningful research direction for clinical scenarios in which patient data is continuously generated. Second, for the diagnosis task, both our model and baselines perform poorly, which needs to be improved in the future.

VI. Conclusion

In this work, we demonstrate the effectiveness of multimodal multitask learning on the task of medical prediction. Our UniMed leverages multimodal EHR data with both structured and unstructured to handle multiple medical prediction tasks simultaneously. UniMed achieves strong performance in each task through the encoder-decoder framework. In addition, experiments show that UniMed learns to adjust the attention weights of input modalitie according to the task and exploits the time-progression correlation between tasks to improve prediction performance. Our UniMed makes a step towards building general-purpose intelligence medical. In the future, we will investigate more medical data from other modalities to UniMed, such as images, waveform data, and omics data.

Acknowledgment

This work was supported by NSFC Grants U19A2067; Science Foundation for Distinguished Young Scholars of Hunan Province (2020JJ2009); National Key R&D Program of China 2017YFB0202602, 2018YFC0910405, 2017YFC1311003, 2016YFC1302500; Science Foundation of Changsha Z202069420652, kq2004010; JZ20195242029, JH20199142034; The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics and Peng Cheng Lab.

References

[1] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] S. Wang, M. B. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann, "Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii," in *Proceedings of the ACM conference on health, inference, and learning*, 2020, pp. 222–235.

[4] M. Sotoodeh and J. C. Ho, "Improving length of stay prediction using a hidden markov model," *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 425, 2019.

[5] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens, "Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1921–1934, 2020.

[6] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, vol. 56. PMLR, 18–19 Aug 2016, pp. 301–318.

[7] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[8] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[9] Z. Qiao, X. Wu, S. Ge, and W. Fan, "Mnn: multimodal attentional neural networks for diagnosis prediction," *Extraction*, vol. 1, p. A1, 2019.

[10] B. Yang and L. Wu, "How to leverage the multimodal EHR data for better medical prediction?" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 4029–4038. [Online]. Available: https://aclanthology.org/2021.emnlp-main.329

[11] Z. Xu, D. R. So, and A. M. Dai, "Mufasa: Multimodal fusion architecture search for electronic health records," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 532–10 540.

[12] V. Parcha, R. Kalra, S. P. Bhatt, L. Berra, G. Arora, and P. Arora, "Trends and geographic variation in acute respiratory failure and ards mortality in the united states," *Chest*, vol. 159, no. 4, pp. 1460–1472, 2021.

[13] Y. Lu, H. Guo, X. Chen, and Q. Zhang, "Association between lactate/albumin ratio and all-cause mortality in patients with acute respiratory failure: A retrospective analysis," *Plos one*, vol. 16, no. 8, p. e0255744, 2021.

[14] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[15] S. A. Franca, C. T. Junior, A. L. D. Hovnanian, A. L. P. Albuquerque, E. R. Borges, V. R. Pizzo, and C. R. R. Carvalho, "The epidemiology of acute respiratory failure in hospitalized patients: a brazilian prospective cohort study," *Journal of critical care*, vol. 26, no. 3, pp. 330–e1, 2011.

[16] G. Strezoski, N. v. Noord, and M. Worring, "Many task learning with task routing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1375–1384.

[17] V. Sanh, T. Wolf, and S. Ruder, "A hierarchical multi-task approach for learning embeddings from semantic tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6949–6956.

[18] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1439–1449.

[19] H. Suresh, J. J. Gong, and J. V. Guttag, "Learning tasks for multitask learning: Heterogenous patient populations in the icu," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 802–810.

[20] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[21] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Multi-task prediction of clinical outcomes in the intensive care unit using flexible multimodal transformers," *arXiv preprint arXiv:2111.05431*, 2021.

[22] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://aclanthology.org/W19-1909

[23] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[24] Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, and K. Rahimi, "Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," *arXiv preprint arXiv:2106.11360*, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[29] D. F. Gaieski and M. Mikkelsen, "Definition, classification, etiology, and pathophysiology of shock in adults," *UpToDate, Waltham, MA. Accesed*, vol. 8, p. 17, 2016.

[30] V. N. Slee, "The international classification of diseases: ninth revision (icd-9)," pp. 424–426, 1978.

[31] S. Robertson, "Nlp from scratch: Translation with a sequence to sequence network and attention," https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html?highlight=attention.