How to Leverage Multimodal EHR Data for Better Medical Predictions?

Bo Yang
Sun Yat-sen University
yangb65@mail2.sysu.edu.cn

Lijun Wu*
Microsoft Research Asia
lijuwu@microsoft.com

Abstract

Healthcare is becoming a more and more important research topic recently. With the growing data in the healthcare domain, it offers a great opportunity for deep learning to improve the quality of medical service. However, the complexity of electronic health records (EHR) data is a challenge for the application of deep learning. Specifically, the data produced in the hospital admissions are monitored by the EHR system, which includes structured data like daily body temperature, and unstructured data like free text and laboratory measurements. Although there are some preprocessing frameworks proposed for specific EHR data, the clinical notes that contain significant clinical value are beyond the realm of their consideration. Besides, whether these different data from various views are all beneficial to the medical tasks and how to best utilize these data remain unclear. Therefore, in this paper, we first extract the accompanying clinical notes from EHR and propose a method to integrate these data, we also comprehensively study the different models and the data leverage methods for better medical task prediction. The results on two medical prediction tasks show that our fused model with different data outperforms the state-of-the-art method that without clinical notes, which illustrates the importance of our fusion method and the value of clinical note features. Our code is available at https://github.com/emnlp-mimic/mimic.

1 Introduction

Under the serious struggle of the COVID-19, the healthcare research domain has attracted more and more attention nowadays. With the improvement of information technology, many hospitals have begun to use EHR (Electronic Health Record) systems to monitor all the data produced during the entire hospital admission. The large amount of data generated in this process offers an opportunity for deep learning technology to improve healthcare, such as diagnoses prediction (Choi et al., 2016), medication recommendation (Shang et al., 2019), mortality prediction (Tang et al., 2020), and readmission prediction (Huang et al., 2019). However, comparing to common academic datasets, such as ImageNet (Deng et al., 2009) and WMT (Macháček and Bojar, 2014), real-world EHR data is longitudinal, heterogeneous, and multimodal, which proposes big challenges to leverage the information included in it.

ICU Stay ID	25130					
Time-Invariant Data	*					
Time-Series Data (Discrete Events)	** *	*		** **		* *
Time-Series Data (Continuous Events)						
Clinical Notes			*			*
Hours	1	2	3	4	5	6

Figure 1: An example of an ICU stay's record within 6 hours. It contains 3 modalities, including time-invariant data, time-series data, and clinical notes. The time-series data can be further split into discrete events and continuous events.

To further understand the complexity of real-world EHR data, we depict the data in Figure 1. The data can be split into three modalities: (1) Time-invariant data, such as age, sex of a patient. Usually, it will not change during the hospital admission. (2) Time-series data, such as vital signs and laboratory measurements. These data share the same property that they change over time and the distribution is nonuniform in time. For instance, vital signs like heart rate and blood pressure are recorded continuously for hours or days, while laboratory measurements such as blood test is a discrete event happened in some time during the admission. (3) Clinical notes, are unstructured free text and normally sparser than the time-series data. More importantly, these notes are full of abbreviations, jargon, and unusual grammatical structures, which are hard for non-professionals to read and

* Corresponding Author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4029–4038
November 7–11, 2021.

understand.

Given the uneven distribution of time-series data, researchers are mostly focused on this data. For example, both Wang et al. (2020) and Tang et al. (2020) extract the raw time-series data into usable hourly features, performing essential operations such as unit conversion, missingness reduction, and outlier handling. However, the clinical notes that contain significant clinical value (Boag et al., 2018) are beyond the realm of these preprocessing pipelines' consideration. Due to the importance of clinical notes, it is necessary to combine them with other data sources for the integrity of clinical features. Since the common pre-trained language models such as BERT (Devlin et al., 2019) do not consider the specific complexity of clinical notes, we apply the ClinicalBERT (Huang et al., 2019) that are pre-trained on clinical notes for handling the notes data in this paper.

Though the above modalities have strong potential in deep learning, modeling the joint representation is nontrivial, as naively adding more features could result in worse performances (Ramachandram and Taylor, 2017). In this work, with the inspiration from Multimodal Adaptation Gate (MAG) (Rahman et al., 2020), we use attention gate for the fusion of the aforementioned three modalities. The core idea behind is to adjust the representation of one modality with a displacement vector derived from the other modalities. For simplicity, we refer to the modality being adjusted as the main modality and the other modalities as auxiliary modalities. The question is, *which modality should be the main modality?* The answer may be clear in multimodal sentiment analysis since language/text always contains the richest sentiment information. However, for medical prediction tasks, such as acute respiratory failure (ARF), we doubt that text/notes should be the main modality since it happens beyond the physician's expectation. Therefore, we conduct experiments on two tasks, diagnoses prediction and acute respiratory failure (ARF) prediction, and we give comprehensive explorations of the fusion strategy. The results show that our model with clinical notes outperforms the models without them, which illustrates the importance of the clinical notes and the effectiveness of our fusion method.

The contributions of this paper can be summarized as follows:

- We first propose to jointly modeling the different data sources extracted from preprocessing pipeline and the clinical notes for improving the medical predictions.
- We propose a fusion method to integrate the time-invariant data, time-series data, and clinical notes with a large pre-trained model.
- Empirical evidence demonstrates the superiority of our fusion model over the traditional models with only the pipeline data as input, which also proves the value of clinical notes.

2 Related Work

We conduct our work based on the data extracted from MIMIC-III (Johnson et al., 2016), a real-world EHR database comprising information relating to patients admitted to intensive care unit (ICU). Recently, there has been surge of methods which apply deep learning on these tabular domain data. Yoon et al. (2020) propose a self- and semi-supervised learning frameworks for value imputation and data augmentation in tabular domain data. Wang et al. (2021) propose an improved Deep & Cross Network (DCN) to learn explicit feature interactions.

Given the complexity of EHR data, Wang et al. (2020) propose a pipeline to transform the raw MIMIC-III data into usable hourly time-series data. To break through the limitation of a specific dataset, Tang et al. (2020) propose a systematic preprocessing technique named FIDDLE for EHR data. We use the data extracted by FIDDLE in this paper.

For single modality tasks, medical codes are commonly extracted and input to RNN for diagnoses prediction (Choi et al., 2016) or medication prediction (Shang et al., 2018, 2019). Recently, since the thrive of multimodal machine learning, researchers have begun to leverage the multimodal nature of EHR data to improve prediction performance (Shin et al., 2019). Qiao et al. (2019) improve diagnoses prediction by combining medical codes and clinical notes through a multimodal attentional neural network. Xu et al. (2018) propose a recurrent attentive model to fuse continuous patient monitoring data, such as electrocardiogram (ECG), and discrete clinical events for predicting length of ICU stay. Comparing to the handcrafted models, Xu et al. (2021) propose a neural architecture search (NAS) method to simultaneously search across multimodal fusion strategies and modality-specific architectures for diagnoses prediction.

Our model builds upon ClinicalBERT (Huang et al., 2019), which has the same architecture with BERT (Devlin et al., 2019). Similarly, ClinicalBERT is pre-trained on the clinical notes of MIMIC-III (Johnson et al., 2016) with two unsupervised tasks, masked language modeling (MLM) and next sentence prediction (NSP). The method of integrating multimodal information into large pre-trained transformers like BERT has also been explored in Rahman et al. (2020). Only that (Rahman et al., 2020) apply attention gate at word-level to combine a lexical input vector with its visual and audio accompaniments for sentiment prediction. Differently, we first pass the features to subnetworks and then fuse the outputs using the attention gate for diagnoses and ARF prediction.

3 Our Method

Our method consists of three logical parts: encoding, fusion, and prediction. In this section, we will describe these parts in detail. For clarity, we first define the data notations used in our method here. The data are split into two categories according to the method used to extract it. The data extracted from preprocessing pipeline framework such as Tang et al. (2020) is split into time-invariant data and time-series data. Given the batch size by B, the time-invariant data can be represented as $\mathbf{I}_{ti} \in \mathcal{R}^{B \times D_1}$, where D_1 is the dimension of time-invariant feature. Similarly, the time-series data is denoted as $\mathbf{I}_{ts} \in \mathcal{R}^{B \times L \times D_2}$, where L represents the length of the ICU stay counted by hours and D_2 is the dimension of the time-series data. For clinical notes, wordpiece (Wu et al., 2016) is applied to tokenize and transform them into token ids. Given the length of the notes by D_3, the token ids are represented as $\mathbf{I}_{nt} \in \mathcal{R}^{B \times D_3}$.

3.1 Encoding

We use different encoders for each modality:

Time-invariant Encoding: Given that time-invariant data contains simple and fixed information of a patient, such as age, sex, and ethnicity, we believe a fully-connected network with `ReLU` activation is enough to encode these information, that is $\mathbf{E}_{ti} = \texttt{ReLU}(\texttt{Linear}(\mathbf{I}_{ti}))$, where $\mathbf{E}_{ti} \in \mathcal{R}^{B \times D'_1}$, and D'_1 is the dimension of the encoded feature.

Time-series Encoding: Given that time-series data consists of hourly features including vital signs, laboratory measurements, and medications, models with the ability to handle temporal sequences are preferred for encoding them. In this work, we use four different encoders, each of which is fused with ClinicalBERT (introduced later) to create a baseline model.

The encoders can be split into two groups according to the different modeling functions and the time they are proposed. The first group contains Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Convolutional Neural Networks (CNN) (LeCun et al., 1998). We choose them because they achieve the best performance on pipeline data in Tang et al. (2020). The second group contains Star-Transformer (Guo et al., 2019) and Transformer encoder (Vaswani et al., 2017). We choose the Transformer encoder because it can learn the representation by jointly conditioned on both left and the right context, and Star-Transformer reduces the complexity of Transformer to linear while preserving the capacity to capture both local composition and long-range dependency.

Formally, the computations of these encoders are as follows:

$$\begin{aligned} \mathbf{E}'_{ts} &= \texttt{ENC}(\mathbf{I}_{ts}), \\ \mathbf{E}_{ts} &= \texttt{ReLU}(\texttt{Linear}(\mathbf{E}'_{ts})), \end{aligned} \tag{1}$$

where $\mathbf{E}'_{ts} \in \mathcal{R}^{B \times L'_2}$ and $\mathbf{E}_{ts} \in \mathcal{R}^{B \times D'_2}$, L'_2 is the hidden size and D'_2 is the number of neurons. `ENC` means encoder, corresponding to the aforementioned four encoders: LSTM, CNN, Transformer encoder and Star-Transformer. The computation of Transformer encoder is a little different from that in Equation (1), where $\mathbf{E}'_{ts}$ is not passed to the `Linear` that follows and directly used as encoded representation. Noting that we always use the hidden state of the last layer if there are multiple layers in LSTM.

Clinical Notes Encoding: As mentioned in Section 1, we use the pre-trained ClinicalBERT to encode the clinical notes. When training for specific tasks, the entire pre-trained model will be fine-tuned with the other encoders for better adaptation. Denoting the encoded feature as $\mathbf{E}_{nt} \in \mathcal{R}^{B \times D'_3}$, then $\mathbf{E}_{nt} = \texttt{ClinicalBERT}(\mathbf{I}_{nt})$.

3.2 Fusion

Inspired by the Multimodal Adaptation Gate (MAG) (Rahman et al., 2020), we make the fusion of the three modalities with attention gate, which

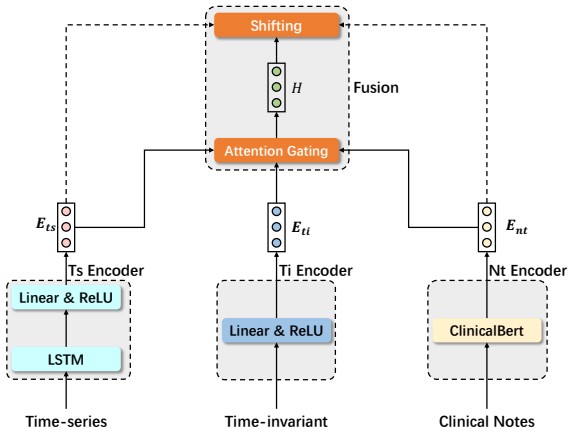

Figure 2: The overall architecture of our proposed model. Ts, Ti and Nt are abbreviations of Time-series, Time-invariant and Clinical Notes. We use dotted line connect *Shifting* module since this connection depends on the task, could be $\mathbf{E}_{ts}$ or $\mathbf{E}_{nt}$.

can be observed in Figure 2. The implementation of MAG is first studied in Wang et al. (2019), where a displacement vector $\mathbf{H}$ is computed by cross-modal self-attention between visual/audio and text modalities. This operation is performed at the word level for shifting the word representation in light of non-verbal cues. However, unlike the video data in their work, the multimodal data in our task is inherently asynchronous, which means there is no accompanying modality for each word. Besides, comparing to the above sentiment analysis task, the importance of text (notes) in medical prediction is not that clear. Therefore, we fuse the modalities at the sample level and switch the main modality as needed. Formally, the computation of `Attention Gating` in Figure 2 is as follows:

$$\begin{aligned} g_1 &= \texttt{ReLU}(\texttt{Linear}([\mathbf{E}_{nt}; \mathbf{E}_{ti}])), \\ g_2 &= \texttt{ReLU}(\texttt{Linear}([\mathbf{E}_{nt}; \mathbf{E}_{ts}])), \end{aligned} \tag{2}$$

where $g_1 \in \mathcal{R}$ and $g_2 \in \mathcal{R}$ are two gating values for time-invariant and time-series modalities.

The displacement vector $\mathbf{H}$ is calculated by merging $\mathbf{E}_{ti}$ and $\mathbf{E}_{ts}$ multiplied by their respective gating values:

$$\mathbf{H} = \texttt{Linear}([g_1\mathbf{E}_{ti}; g_2\mathbf{E}_{ts}]), \tag{3}$$

where $\mathbf{H} \in \mathcal{R}^{B \times D'_3}$. Given the main modality by clinical notes, a weighted summation is performed between the main feature $\mathbf{E}_{nt}$ and the displacement vector $\mathbf{H}$ to create a multimodal representation $\mathbf{M}$:

$$\begin{aligned} \mathbf{M} &= \mathbf{E}_{nt} + \alpha\mathbf{H}, \\ \alpha &= \texttt{min}(\frac{||\mathbf{E}_{nt}||_2}{||\mathbf{H}||_2}\beta, 1), \end{aligned} \tag{4}$$

where β is a randomly initialized hyper-parameter training with the model. $||\mathbf{E}_{nt}||_2$ and $||\mathbf{H}||_2$ are the L_2 norm of $\mathbf{E}_{nt}$ and $\mathbf{H}$, respectively. The scaling factor α is used to restrict the effect of the displacement vector $\mathbf{H}$ to a desirable range.

Note that Equation (2) , Equation (3) and Equation (4) take clinical notes as the main modality, the switch of the main modality can be performed by switching the position of $\mathbf{E}_{nt}$ with other expected representation such as $\mathbf{E}_{ts}$ in these equations.

3.3 Prediction

Following Choi et al. (2016), we use the *softmax* layer to produce the prediction for the multi-label problems such as diagnoses prediction, that is:

$$\hat{\mathbf{Y}} = \texttt{softmax}(\texttt{Linear}(\mathbf{M})), \tag{5}$$

where $\hat{\mathbf{Y}} \in \mathcal{R}^{B \times N}$ and N is the number of labels. For the binary classification problems such as ARF prediction (introduced later), we apply *sigmoid* layer instead of *softmax*:

$$\hat{\mathbf{y}} = \texttt{sigmoid}(\texttt{Linear}(\mathbf{M})), \tag{6}$$

where $\hat{\mathbf{y}} \in \mathcal{R}^B$.

Following Qiao et al. (2019) and Tang et al. (2020), we use cross-entropy between the ground truth and the prediction to compute the loss for all the ICU admissions in both of the above problems. Given the ground truth of the multi-label problems by $\mathbf{Y}$, the computation is:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \mathbf{Y}_i\texttt{log}(\hat{\mathbf{Y}}_i) + (1-\mathbf{Y}_i)\texttt{log}(1-\hat{\mathbf{Y}}_i), \tag{7}$$

where $\mathcal{L} \in \mathcal{R}$. For binary classification problems, the loss is calculated by replacing the ground truth $\mathbf{Y}$ and prediction $\hat{\mathbf{Y}}$ in Equation (7) to its own.

4 Experiments

In this section, we first introduce the MIMIC-III dataset. Then we describe the tasks and the metrics used to evaluate the performance. After that, we introduce the baseline models adopted in this paper and the detailed experimental design of this work.

4.1 Dataset

Dataset We use the Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016) dataset. It contains real-world EHR data including vital signs, laboratory measurements, and clinical notes (free text) relating to ICU patients at the Beth Israel Deaconess Medical Center between 2001 and 2012. We focus on the $17,710$ patients ($23,620$ ICU visits) recorded using the iMDSoft MetaVision system from 2008 to 2012 since they represent more up-to-date practices. In our experiments, the non-text features within 48 / 12 hours are extracted using FIDDLE (Tang et al., 2020). They are randomly split into train, validation, and test sets in a $7 : 1.5 : 1.5$ ratio. The text features within 48 / 12 hours are produced by gathering the latest notes of each category into one document and tokenized by WordPiece (Wu et al., 2016). Given the time limitation and modality requirement, we exclude the patients who stay in ICU less than 48 / 12 hours and the patients with incorrect notes or without notes. After data preprocessing, there remain $10,210$ / $14,174$ samples.

4.2 Prediction Tasks and Metrics

For each ICU visit, we use the EHR data recorded for the following prediction tasks:

Diagnoses: Predicting the diagnoses by using the data produced within 48 hours from the start of this ICU admission. This is a multi-label problem since each visit may relate to multiple diseases. The label is produced by transforming the corresponding International Classification of Diseases, 9th Revision (ICD-9) diagnosis code (Slee, 1978) into a multi-hot vector. Following Qiao et al. (2019), we extract the top-3 digits of ICD-9 in the ICD-9 definition table of MIMIC-III, yielding $1,042$ disease groups. We use Top-k recall (Choi et al., 2016) to evaluate this task since it mimics the behavior of doctors conducting a differential diagnosis, where doctors list the most probable diagnoses and treat patients accordingly to identify the patient status. In our experiments, we separately set k to be 10, 20, and 30.

ARF: Predicting whether the patient will fall into acute respiratory failure (ARF) by using the data produced within 12 hours from the start of this ICU admission. This is a binary classification problem. Following Tang et al. (2020), we use AUROC (Area Under the Receiver Operating Characteristic curve) and AUPR (Area Under the Precision-Recall curve) to evaluate the performance.

Noting that our diagnoses task is a brand new task that is different from the diagnoses tasks introduced in Section 2. In Choi et al. (2016); Qiao et al. (2019), the task is to predict the diagnoses of the next visit by using the previous ICU admission data. In Xu et al. (2021), the task is to predict the diagnoses of the current visit by using all the data generated during this ICU admission. We argue that in clinical practice, the earlier a diagnose is made, the more valuable it is. Thus, we extract the first 48 / 12 hours data in this work instead of the entire admission data for diagnoses prediction. Besides, the input data of our model is also distinct from them, for example, the medical codes are very important components in their input data, while they are not included in our input data since they are not yet generated during the time limitation. Given that these models are designed specifically for their input data, we cannot compare the performance of them with our model. Therefore, we exclude them from our baselines.

4.3 Baseline Models

For ARF task, Tang et al. (2020) have explored several traditional machine learning methods including logistic regression and random forest, as well as some deep learning methods like LSTM and convolutional neural network (CNN), we denote them as *F-LR*, *F-RF*, *F-Lstm* and *F-Cnn* respectively. All of them are included in our baselines. We also introduce some models proposed more recently such as Transformer (Vaswani et al., 2017) encoder and Star-Transformer (Guo et al., 2019) for the single modality baselines. Besides, we combine each of the above encoder with ClinicalBERT (Huang et al., 2019) to create 4 multimodal models, *LstmBert*, *CnnBert*, *StarBert* and *EncoderBert*, where *Lstm*, *Cnn*, *Star*, *Encoder* and *Bert* represent LSTM, CNN, Star-Transformer, Transformer encoder, and ClinicalBERT, respectively. Each model name can be split into two modules, where the first one suggests the main modality in the fusion. For example, *LstmBert* means time-series data is the main modality since *Lstm* is the encoder of time-series data. In that way, the main modality of all the 4 models is time-series data. To study the effect of main modality, we switch the main modality of the above 4 multimodal models to create another 4 models: *BertLstm*, *BertCnn*,

Task	ARF		Diagnoses		
Metric	AUROC	AUPR	Recall@30	Recall@20	Recall@10
F-LR (Tang et al., 2020)	0.757	0.291		-	
F-RF (Tang et al., 2020)	0.760	0.317		-	
F-Lstm (Tang et al., 2020)	0.771	0.326	0.558	0.461	0.312
F-Cnn (Tang et al., 2020)	0.768	0.294	0.553	0.458	0.312
BertLstm	0.772	0.278	0.464	0.376	0.243
LstmBert	**0.792**	**0.350**	0.553	0.457	0.305
BertCnn	0.778	0.348	0.465	0.376	0.243
CnnBert	0.753	0.304	0.521	0.426	0.285
BertStar	0.687	0.249	0.560	0.465	0.314
StarBert	0.687	0.262	0.559	0.465	0.313
BertEncoder	0.730	0.294	**0.587**	**0.490**	**0.334**
EncoderBert	0.695	0.276	0.547	0.456	0.314

Table 1: The results of fusion models on ARF and Diagnoses prediction. The best results are highlighted in bold. The prefix F in the first part means the fusion of time-invariant and time-series data. *LR* denotes logistic regression and *RF* denotes random forest. The meaning of the other models are stated in Section 4.3.

BertStar and *BertEncoder*. The main modality in these models is switched to clinical notes. The computation of fusion and switch is introduced in Section 3.2. We use all the models aforementioned as our baselines.

4.4 Experimental Design

For the selection of model parameters, we first refer to the papers that performing similar tasks or applying similar models for choosing reasonable ranges of each parameter. Then we further screen these parameters to reduce the cost and use grid search to find the parameter combination that performs best on the validation set. Specifically, we train all the models in this paper using Adam (Kingma and Ba, 2014) with a learning rate of $1e-4$. The dropout of each model is set to 0.1. For ClinicalBERT, we adopt the default configuration of the BERT$_{\text{BASE}}$ model and load the pre-trained parameters for fine-tuning. The dimension of encoded time-invariant data is set to 64 for all the models, which corresponding to D_1' in Section 3.1. For the ARF task, the model that achieves the best results is *LstmBert*, where we set the hidden size L_2' to 512 and the number of neurons D_2' to 128. We use a single *Lstm* layer. The number of parameters in *LstmBert* is 120M. For the Diagnoses task, the best model is *BertEncoder*, where the hidden size and number of layers of the *Encoder* are set to 1024 and 3, respectively. The number of parameters in *BertEncoder* is 150M.

5 Results and Discussion

In this section, we first evaluate the performance of all the fused models on the ARF and Diagnoses tasks, and discuss the effect of the main modality on each of them. Then we perform an ablation study on the best model for understanding the influence of individual modules in our method. Finally, we perform experiments by using other fusion strategies to study the effect.

5.1 Results of ARF

The results of ARF are shown on the left side of Table 1, where the models are split into three parts. In the first part, we use the results published in Tang et al. (2020), where the input feature of each model is created by concatenating the time-invariant and time-series data, we use the prefix *F* to represent this fusion. The detailed description is in Section 5.4. These models represent the state-of-the-art models without clinical notes. In the second part, we fuse the ClinicalBERT with the classical deep learning models LSTM and CNN. In the third part, we fuse the ClinicalBERT with the recently proposed model Star-Transformer and Transformer Encoder. First, the best performance is achieved by *LstmBert*, and comparing part 1 with part 2, we can see that most models in part 2 outperform the models in part 1, which illustrates the value of the clinical notes and also the effectiveness of the fusion method. Second, we also observe that the models in part 1 and part 2 generally outperforms

the models in part 3, which illustrates that the time-series encoder (e.g., *Lstm*, *Star*) can significantly influence the performance of the fusion models, even makes the performance inferior (e.g., *Encoder*, *Star*) to the models without clinical notes. The results further illustrate the argument that adding more features naively could result in worse performances, even though conditioned on the same fusion method. Besides, comparing to other fusion pairs, the performance of *BertStar* and *StarBert* are close to each other.

5.2 Results of Diagnoses

The performance of fusion models on Diagnoses prediction is shown on the right side of Table 1. Comparing the results of part 3 with part 1, we find that the fusion models *BertStar*, *StarBert* and *BertEncoder* outperform the best results in part 1, which further demonstrates the effectiveness of our method. Comparing to the results of ARF, we observe an opposite trend in this task. Specifically, the results in part 3 generally outperform those in part 1 and part 2. This trend suggests that the time-series encoder should be chosen carefully for different tasks since the superiority of one encoder is not guaranteed to generalize to other tasks. Besides, similar to ARF prediction, the performance of *BertStar* and *StarBert* are also very close in this task, which suggests that ClinicalBERT and Star-Transformer in our fusion method have nearly equal status. We also notice that the models that achieve the best results on the two tasks have different main modalities. For ARF, it is *LstmBert* with time-series data as the main modality. For Diagnoses, it is *BertEncoder* with clinical notes as the main modality. It demonstrates the distinct significance of each modality in different tasks. We will further discuss it in the ablation study.

5.3 Ablation Study

In this section, we gradually remove the components of the fusion models to explore their effect on the performance. The results are shown in Table 2.

First, we use only the time-invariant data for the prediction of ARF and diagnoses. The results are given by *Ti*. There is a large gap between them and the others since the information included in time-invariant data is not as much as in the other two modalities.

Then, we use only the time-series data for the prediction of ARF and diagnoses. The results are given by *Lstm*, *Star*, *Encoder* and *Cnn*, which corresponding to the four encoders applied in time-series encoding. Then we remove the time-series encoder and make predictions only with the clinical notes. The results are given by *Bert*.

For the ARF task, we observe that all the four time-series encoders outperform *Bert* by a large margin. This consistent superiority illustrates that time-series data is more effective than clinical notes in this task. This is intuitive since ARF (Acute Respiratory Failure) is an emergency, the notes taken by physicians and nurses are unlikely to predict it more accurate than the real-time time-series data like vital signs. The best results of single modality models are achieved by *Lstm*. Given the importance of time-series data and the performance of *Lstm*, it is reasonable to infer that *LstmBert* would be the best fusion model. This inference has been authenticated by the experimental results in Table 1. We also introduce the results of *LstmBert* as *Fusion* of ARF in Table 2 for comparison. Comparing to *Lstm*, *LstmBert* achieves significant improvements on AUROC and AUPR, which demonstrates the effectiveness of the fusion method and the value of note features.

For the Diagnoses task, contrary to ARF, *Bert* outperforms the four time-series encoders significantly. The opposite trend suggests that clinical notes should be the main modality in this task. Comparing to time-series data like blood tests, clinical notes can provide an accurate description of clinical symptoms. In terms of disease diagnoses, we believe the clinical symptoms are the most important and direct evidence for the physicians' opinion, while the laboratory measurements are commonly used for auxiliary diagnoses, such as confirming the opinion or excluding other diseases. Therefore, it is reasonable for the clinical notes to possess the dominant position. In this case, the importance of the time-series encoders is relatively reduced, which explains why the best encoder is *Star* while the best fusion model is *BertEncoder* (refer to *Fusion* of Diagnoses in Table 2). In this task, the fusion model *BertEncoder* still outperforms the best single modality model *Bert*, which illustrates the availability of time-series features.

5.4 Other Fusion Strategies

We further explore two strategies for the fusion of the three modalities. These fusion strategies have been studied in several influential works on multimodal sentiment analysis (MSA). In MSA, the

Task	ARF		Diagnoses		
Metric	AUROC	AUPR	Recall@30	Recall@20	Recall@10
Ti	0.600	0.151	0.513	0.420	0.278
Lstm	0.772	0.336	0.547	0.451	0.310
Cnn	0.767	0.327	0.549	0.455	0.313
Star	0.768	0.326	0.558	0.466	0.320
Encoder	0.749	0.284	0.548	0.450	0.304
Bert	0.692	0.255	0.577	0.479	0.330
Fusion	**0.792**	**0.350**	**0.587**	**0.490**	**0.334**

Table 2: The results of ablation study on ARF and Diagnoses prediction. The best results are highlighted in bold. For simplicity, we use *Fusion* represents the best fusion model of each task and *Ti* is the time-invariant encoder.

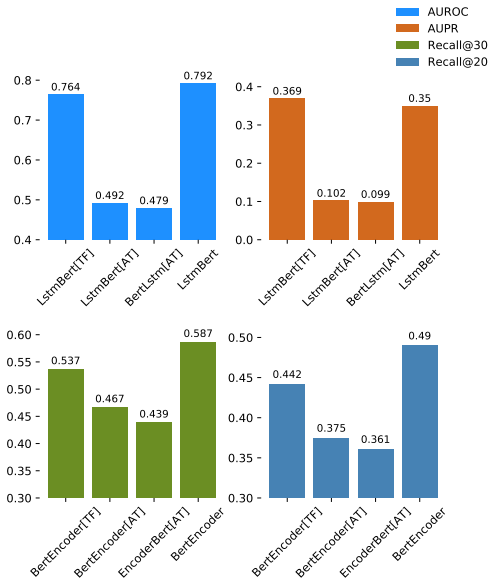

Figure 3: The results of other fusion strategies on ARF and Diagnoses prediction. `TF` means using the tensor fusion method to merge the two modules. `AT` means using an attention mechanism to merge the two modules. Since attention is computed asymmetrically like MAG (Rahman et al., 2020), we also use the first module as the main modality like Table 1.

multimodal data including visual, audio, and text modalities are extracted from a video, which means they are synchronous and theoretically have equal status. However, in our task, the time-invariant data such as age, sex is not as important as the time-series data and clinical notes in terms of the amount of information it contains. Therefore, we adjust the origin strategy to our task by splitting the fusion process into two stages.

Following Tang et al. (2020), we perform an early fusion on time-invariant data and time-series data in the first state. Specifically, given the input feature of the two modalities by $\mathbf{I}_{ti} \in \mathcal{R}^{B\times D_1}$ and $\mathbf{I}_{ts} \in \mathcal{R}^{B\times L\times D_2}$, we first extend $\mathbf{I}_{ti}$ to $\mathbf{I}'_{ti}$ by repeating the feature vector L times, thus $\mathbf{I}'_{ti} \in \mathcal{R}^{B\times L\times D_1}$. After that, we concatenate $\mathbf{I}'_{ti}$ with the time-series feature $\mathbf{I}_{ts}$ for the fusion, that is $\mathbf{I}_t = [\mathbf{I}'_{ti}; \mathbf{I}_{ts}]$, where $\mathbf{I}_t \in \mathcal{R}^{B\times L\times D_t}$ and $D_t = D_1 + D_2$. Finally, the fused vector $\mathbf{I}_t$ is passed to the encoders for the encoded vector $\mathbf{E}_t$ just like the **Time-series Encoding** in Section 3.1. Besides, similar to the **Clinical Notes Encoding** in Section 3.1, we also use ClinicalBERT to encode the clinical notes in this section.

Given the encoded text vector by $\mathbf{E}_{nt}$, we focus on the fusion of $\mathbf{E}_{nt}$ and $\mathbf{E}_t$ in the second stage. We provide two fusion strategies for it. The first one is tensor fusion (Zadeh et al., 2017; Liu et al., 2018), where the fusion is performed by an outer product of the encoded representations of different modalities. The second one is attention fusion (Tsai et al., 2019). The core idea of this fusion is to attend one modality to another and vice versa.

We adjust the best fusion models in the ARF and Diagnoses tasks to both of the fusion strategies, meaning that the fusion method MAG is replaced by the two strategies in these models. The results are shown in Figure 3. This figure is split into two parts. The upper part is the results of ARF prediction and the lower part is the results of Diagnoses prediction. Since the computation of attention is asymmetric, we also switch the main modality for this fusion method. In this case, the main modality corresponding to the Key and Value.

For the ARF task, we observe that the model with tensor fusion significantly outperforms the models with attention fusion. We attribute this result to the asynchronous input modalities. Since they are not strictly synchronized, and even convey different meanings, it is not reasonable to attend

one modality to another. This is also observed in Diagnoses prediction, which further tests the hypothesis. Besides, the results show that *LstmBert[AT]* outperforms *BertLstm[AT]*, which illustrates that the superiority of time-series data is preserved in attention fusion for ARF task. Similarly, the superiority of *BertEncoder[AT]* also demonstrates the dominant position of clinical notes in the Diagnoses task. Though these fusion strategies provide us new perspectives about how to integrate the modalities, both of them are inferior to the fusion method proposed in this paper.

6 Conclusion

In this paper, we propose to integrate the data extracted from the preprocessing pipeline and the accompanying clinical notes for better medical prediction. Enlightened by the MAG method, we propose a fusion method for our tasks and explore four different encoders to study the effect. Besides, to understand the importance of each modality, we switch the main modality in the fusion models and find that time-series data and clinical notes are the main modalities of the ARF task and Diagnoses task respectively. Finally, we investigate other fusion strategies and the results show that our fusion method achieves state-of-the-art performance. In the future, we will investigate the tabular data related methods stated in Section 2 to see if they can improve the performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China [grant number U1711263].

References

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1315–1325, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014. Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. Mnn: Multimodal attentional neural networks for diagnosis prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5937–5943. International Joint Conferences on Artificial Intelligence Organization.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

Junyuan Shang, Shenda Hong, Yuxi Zhou, Meng Wu, and Hongyan Li. 2018. Knowledge guided multi-instance multi-label learning via neural networks in medicines prediction. In *Asian Conference on Machine Learning*, pages 831–846. PMLR.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133.

Bonggun Shin, Julien Hogan, Andrew B Adams, Raymond J Lynch, Rachel E Patzer, and Jinho D Choi. 2019. Multimodal ensemble approach to incorporate various types of clinical notes for predicting readmission. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE.

Vergil N Slee. 1978. The international classification of diseases: ninth revision (icd-9).

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. 2020. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ruoxi Wang, Rakesh Shivanna, Derek Zhiyuan Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: improved deep & cross network and practical lessons for web-scale learning to rank systems. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1785–1797. ACM / IW3C2.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 222–235.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 2018. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2565–2573.

Zhen Xu, David R So, and Andrew M Dai. 2021. Mufasa: Multimodal fusion architecture search for electronic health records. *arXiv preprint arXiv:2102.02340*.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. 2020. Vime: Extending the success of self- and semi-supervised learning to tabular domain. In *Advances in Neural Information Processing Systems*, volume 33, pages 11033–11043. Curran Associates, Inc.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.