A Multimodal Transformer: Fusing Clinical Notes with Structured EHR Data for Interpretable In-Hospital Mortality Prediction

Weimin Lyu, M.S.[1], Xinyu Dong, M.S. [1], Rachel Wong, M.D., M.B.A., M.P.H. [1], Songzhu Zheng, M.S. [1], Kayley Abell-Hart, B.S. [1], Fusheng Wang, Ph.D. [1,*], Chao Chen, Ph.D. [1,*]
[1]Stony Brook University, Stony Brook, NY, USA

Abstract

Deep-learning-based clinical decision support using structured electronic health records (EHR) has been an active research area for predicting risks of mortality and diseases. Meanwhile, large amounts of narrative clinical notes provide complementary information, but are often not integrated into predictive models. In this paper, we provide a novel multimodal transformer to fuse clinical notes and structured EHR data for better prediction of in-hospital mortality. To improve interpretability, we propose an integrated gradients (IG) method to select important words in clinical notes and discover the critical structured EHR features with Shapley values. These important words and clinical features are visualized to assist with interpretation of the prediction outcomes. We also investigate the significance of domain adaptive pretraining and task adaptive fine-tuning on the Clinical BERT, which is used to learn the representations of clinical notes. Experiments demonstrated that our model outperforms other methods (AUCPR: 0.538, AUCROC: 0.877, F1:0.490).

Introduction

Electronic health record (EHR) systems are widely used in the United States[1] and the large amount of EHR data generated provides an opportunity for machine learning based predictive modeling to improve clinical decision support. In particular, deep learning based techniques,[2,3] have been used for prediction of in-hospital mortality,[4,5] diagnoses,[6,7] length of stay,[8] readmission.[9]

EHRs include structured data and clinical notes, which are often unstructured[10]. Structured clinical variables, such as the vital signals (e.g., heart rate, respiration rate, temperature, and blood pressure), can be easily converted to time series data and are well explored by researchers.[11,12,13] For example, Harutyunyan et al.[14] establishes a benchmark on how to pre-process the MIMIC III dataset, and proposed various baselines for different tasks, e.g., Logistic Regression, Random Forest, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). Dong et al.[6] develops a machine learning based opioid overdose prediction method with different clinical variables. Unstructured clinical notes are often in free narrative form, but contain complementary and rich contextual information, such as a patient's symptoms, disease course and treatment[15]. Though the normally pre-trained natural language model Bidirectional Encoder Representations from Transformers (BERT)[16] cannot handle specific clinical notes, there are different variants of BERTs,[14,17,18] that are pretrained on biomedical and clinical data, which can better handle clinical notes. For example, Clinical BERT[19] pretrains the BERT using MIMIC III clinical notes with masked language modeling (MLM) and next sentence prediction (NSP), to predict hospital readmission. BEHRT[20] incorporates age and position information when modeling the clinical notes. BioBERT[15] is pretrained on biomedical notes like PubMed abstracts and PubMed Central full-text articles to significantly improve biomedical text mining task performance. BioRoBERTa[17] points out that in-domain pretraining leads to performance gains. BLURB[21] benchmark is a recent work that released state-of-the-art pretrained and task-specific models for the community. Despite these advances, how to leverage and interpret the information included in unstructured clinical notes remains a challenging problem.

Multimodal fusion is a promising direction to address the aforementioned challenge. However, naively concatenating features from different modalities might result in worse performances.[22] It is challenging to embed data from the structured clinical variables and unstructured clinical notes together because they are two totally different domains. Si et al.[13] provides a comprehensive survey on deep representation learning from single and multiple domains of EHR data. Some works merely extract features from structured and unstructured data separately, and then concatenate the two features.[23] For example, Khadanga et al.[24] extracts clinical notes with convolutional neural networks and incorporates time series data to improve the performance. Deznabi et al.[25] models two modalities with Long short-term memory (LSTM) and BERT. Yang et al.[7] implements Multimodal Adaptation Gate (MAG)[26] techniques to best utilize information from two modalities. Teixeira et al.[27] tested different combinations of several different modalities.

[*] Contact: Fusheng Wang: Fusheng.wang@stonybrook.edu; Chao Chen: chao.chen.1@stonybrook.edu

Huang et al.[28] discuss fusion strategies of structured data and imaging data. However, these methods naively concatenate without considering complexity of modality and time information. While transformers are gaining popularity for use in different domains, there is limited work using transformers on EHR based predictive modeling.

In this paper, we propose a multimodal transformer to fuse time series data from clinical variables with textual information from clinical notes to boost performance of in-hospital mortality prediction. We leverage clinical notes to provide auxiliary information by adjusting the representation from two modalities to a sharable space across different times, then jointly learn the representation from two modalities. Further, we implement the transformer on the time series EHR data to fully consider the information across time, combined with the fine-tuned Clinical BERT model, which is a novel application of EHR feature representation learning. We also show that pre-training of various BERT models results in different prediction ability with regard to clinical tasks, and the BERT models fine-tuned on the specific in-hospital prediction task brings further performance improvements. Furthermore, we extend our fine-tuned Clinical BERT model with the integrated gradients (IG) method to interpret and visualize the important words in clinical notes. Our results demonstrate that by leveraging the clinical notes, our proposed Multimodal Transformer provides highly promising prediction results (with AUCPR: 0.538, AUCROC: 0.877, F1:0.490). *To our best knowledge, our Multimodal Transformer is the first work utilizing the transformer block to fuse clinical notes information and clinical variable information, while including time series EHR data.*

Methods

Study Dataset

We extract inpatient EHR data from the Medical Information Mart for Intensive Care (MIMIC-III) dataset.[29] The clinical variables are pre-processed as time series signals from ICU instruments following Harutyunyan et al.'s[14] benchmark setup. Seventeen clinical variables remained after preprocessing: capillary refill rate; diastolic blood pressure; fraction inspired oxygen; the eye opening, motor response, verbal response, and total value of the Glasgow Coma Scale; glucose; heart rate; height; mean blood pressure; oxygen saturation; respiratory rate; systolic blood pressure; temperature; weight; and pH.

For the clinical notes, similar to the setup from Khadanga et al.[24], we extract notes from the NOTEEVENTS.csv file, and remove all clinical notes that do not have any chart time associated and remove patients that do not have any clinical notes. While Khadanga et al. kept only the first visit for each patient, we treat every visit as a single sample. Therefore, in the following paper, we use 'patient' to indicate 'visit'. After the above data processing, our dataset for in-hospital mortality prediction contains 14068 training samples, 3086 validation samples, and 3107 test samples. Due to this step, our results are not directly comparable to the benchmark from Harutyuanyan et al..[14]

	Type	Variable Name	Value
Clinical Variables	Categorical	Capillary refill rate	1
		Glascow coma scale eye opening	3
		...	...
	Numerical	Diastolic blood pressure	59.1
		Fraction inspired oxygen	0.21
		...	...
Clinical Notes	Text	Notes	chest (portable ap) clip # reason : pt had a left sided picc line placed...

Figure 1. An example of MIMIC III EHR data for ICU patients, containing two modalities: clinical variables and clinical notes. The clinical variables can be further split into categorical and numerical variables, while the clinical notes are domain specific text.

Single Model Embedding

We aim to predict in-hospital mortality with multi-modal EHR data. First, we process two single modalities (clinical notes and clinical variables) separately to obtain the initialized embeddings from the raw data. We introduce Notes Embedding and Variables Embedding to achieve the initialized single modality embedding.

Clinical Notes Embedding. Though BERT models dominate increasing numbers of domains in NLP, BERT-based models do not necessarily offer strong clinical text mining ability with regard to a specific clinical task. Rather, the power of BERT-family models relies on domain adaptive pretraining and task adaptive fine-tuning. Pre-training on

proper clinical biomedical corpora enables the BERT-based model to better learn clinical contextual meaning representations, and fine-tuning on downstream tasks can further boost this ability and establish a specialized Clinical BERT model. To illustrate this point, we compare the representation ability of four different BERT models (Table 1) using only single clinical notes modality, namely BERT,[16] BioBERT,[15] BioRoBERTa,[17] Clinical BERT,[19] pertained on four types of corpora, respectively English Wikipedia / BooksCorpus, PubMed Abstracts / PMC Full-text articles (initialized from BERT), S2ORC,[33] and entire MIMIC III notes (initialized from BioBERT). Detailed results are shown in Table 4.

We select Clinical BERT[19] as our pre-trained language model since it is a more proper domain-specific model trained on all MIMIC-III clinical notes. We further fine-tune the Clinical BERT with the in-hospital mortality prediction task on MIMIC-III, called MIMIC BERT (MBERT), which enables the Clinical BERT to learn better clinical specific contextual embeddings on specific MIMIC data. For each patient, we extract an embedding of the clinical notes for every associated hour to represent the clinical notes data with time information. In the following experiments, we freeze the weights of MBERT when extracting unstructured clinical notes embeddings in multimodal transformer, since the MBERT already preserves a good clinical meaning representation.

Table 1. Four BERT models and their respective corpora used for pretraining. Initialized model indicates the starting point before pretraining.

Pretrained Model	Pretraining Corpora	Initialized Model	Domain
BERT	English Wikipedia, BooksCorpus		General
BioRoBERTa	S2ORC	RoBERTa	Biomedical
BioBERT	PubMed Abstracts, PMC Full-text articles	BERT	Biomedical
Clinical BERT	MIMIC notes	BioBERT	Biomedical

Clinical Variables Embedding. Given that clinical variables contain numerical and categorical data, we apply one-hot encoding to the clinical variables, illustrated in Figure 2. Following Harutyunyan's setup, the 17 clinical variables are embedded to a 76-dimension time series embedding after the one-hot encoding. The categorical variables are converted to one-hot vectors while the numerical variables are converted to a single continuous value.

For a formal mathematical representation, we denote the clinical notes data as $X_{notes} \in \mathcal{R}^{L \times D_1}$, where L represents the length of ICU stay counted by hours, and D_1 represents the maximum length of clinical notes. And we denote the clinical variables data as $X_{ts} \in \mathcal{R}^{L \times D_2}$, where D_2 represents the number of variables. The clinical notes are embedded with Fine-tuned Clinical BERT (MBERT), as $E_{notes} = MBERT(X_{notes})$, and the clinical variables are embedded as $E_{ts} = Variable\_Encoding(X_{ts})$.

Clinical Variables Encoding

Visit ID	Capillary refill rate (One-Hot)		Diastolic blood pressure	Fraction inspired oxygen	Glascow Coma Scale eye opening (One-Hot)								... (Other Variables)
Visit 1	0	1	59.1	0.21	0	0	0	1	0	0	0	0	
Visit 2	1	0	58.7	0.23	1	0	0	0	0	0	0	0	
Visit 3	1	0	46.3	0.35	0	0	1	0	0	0	0	0	

... (Other Visits)

Figure 2. An illustration of Clinical Variables Encoding.

Multimodal Embedding

We introduce a transformer to integrate two different modalities. Specifically, we introduce three encoders inside the transformer block: *Notes Encoder* and *Time Series (TS) Encoder* for clinical notes and clinical variables modalities separately, and *Multimodal (MM) Encoder* to fuse two modalities while projecting them into a shared space:

Encoders. (1) Notes Encoder. Given that the clinical notes embedding E_{notes} is already well presented, we only use a single linear layer to project E_{notes} to a universal space. (2) Time Series (TS) Encoder. Since the clinical variables

embedding contains simple numerical and categorical information, we also use linear layers to project E_{ts} to a universal space. (3) Multimodal (MM) Encoder. We do not simply concatenate the clinical notes and clinical variables because the two modalities are conceptually different. We use a Multimodal Encoder to compact the two different modalities into a universal space before we feed them into the transformer model, so that the information from clinical notes and clinical variables can be jointly learned. The formal mathematical representation is as follows:

$$\begin{cases} I_{notes} = Encoder_{notes}(E_{notes}) \\ I_{ts} = Encoder_{ts}(E_{ts}) \\ I_{MM} = Encoder_{MM}(E_{notes} \oplus E_{ts}) \end{cases}$$

where $\oplus$ denotes concatenate operation, $I_{notes} \in \mathcal{R}^{L \times D_3}$, $I_{ts} \in \mathcal{R}^{L \times D_4}$, $I_{MM} \in \mathcal{R}^{L \times D_5}$ denotes outputs from associate encoders, and D_3, D_4, D_5 represent the corresponding embedding dimension.

Transformer. Our transformer block is the key to handle time series embeddings and integrate knowledge. The transformer is a popular model developed for natural language processing (NLP), and has emerged as a promising tool in other domains. In LSTM, if the time sequence is too long, then when the information is passed to the final timestamp, the model forgets the information in the earlier timestamps. The powerful attention mechanism in the transformer enables the model to better leverage the information from all the timestamps. However, more research is needed to determine how best to apply the transformer in clinical tasks, especially when using multimodal data. We successfully implement the transformer in our study to show the capability of the transformer model.

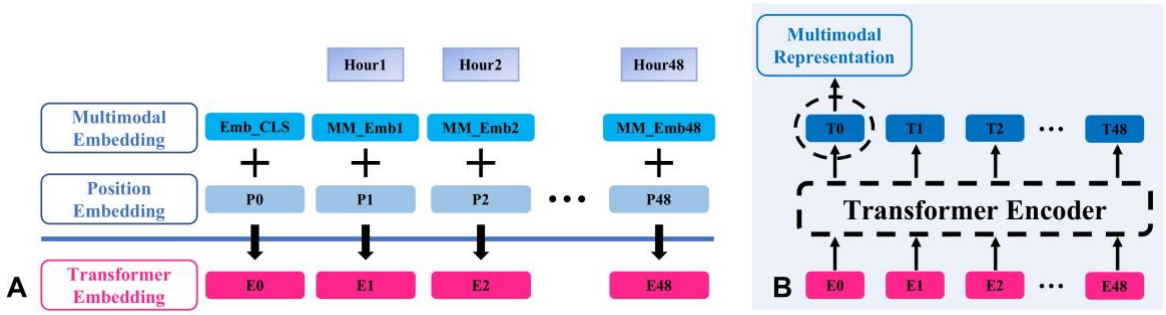


Figure 3. An illustration of the Transformer architecture. A. Adding the position embedding to consider time information. B. Details of the transformer block. The fused transformer embedding is fed into the transformer encoder, and we only select the 'T0' token as the final multimodal representation.

In the NLP context, if one sentence has 48 tokens after tokenization, then the input of the NLP model would be a 48-length sequence. Similarly, in the clinical time series context, we treat each hour as one token. Since we consider the first 48 hours in the ICU, there are 48 'tokens' for one patient. In Figure 3, the multimodal embedding of one patient is shown. The position embedding encodes the time information. In this way, the transformer block is able to consider information from all the time sequences when learning the representations. We use sinusoidal positional embedding in our model. Similar to the NLP techniques, we insert the 'CLS' token at the beginning of the time sequences and use the T0 as the final multimodal representation. Figure 4 illustrates the detailed architecture of the whole transformer block, with a formal mathematical presentation:

$$I_{Multimodal} = Transformer(I_{MM})$$

Then we concatenate the multimodal representations $I_{Multimodal}$ and notes embedding E_{notes} to get the final prediction:

$$Pred = MLP(I_{Multimodal} \oplus E_{notes})$$

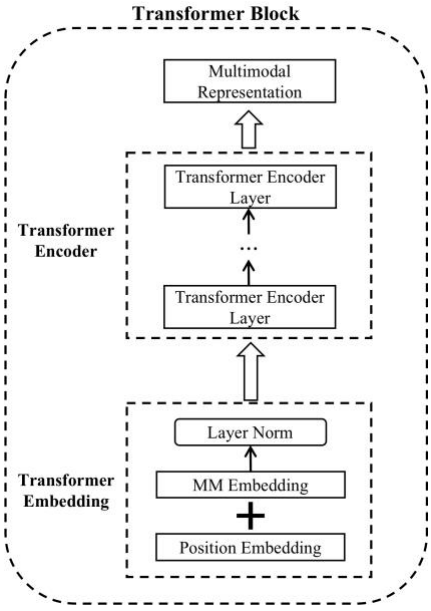


Figure 4. An illustration of the Transformer block.

Overview Architecture

The overall architecture of our Multimodal Transformer is shown in Figure 5.

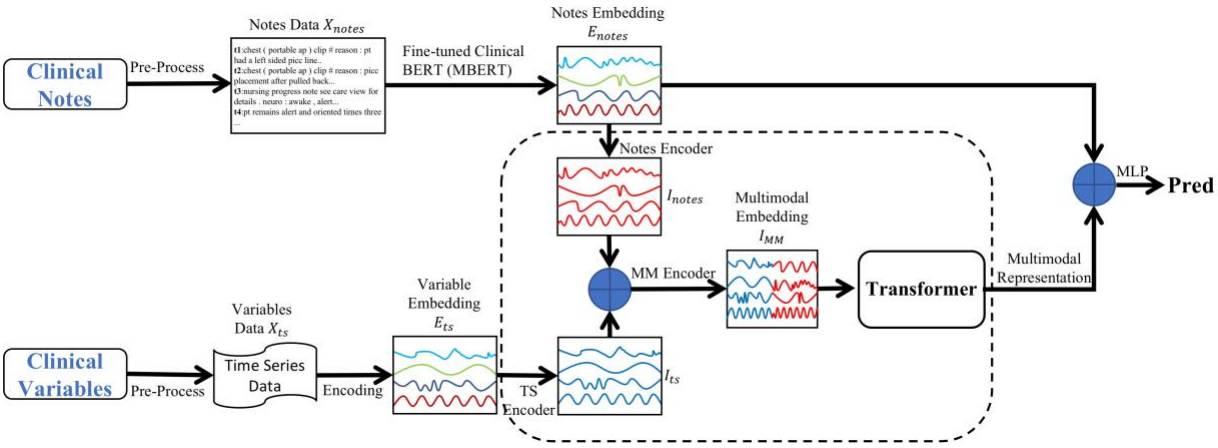


Figure 5. The overview architecture of our proposed Multimodal Transformer.

Implementation. In our experiment, a rectified linear unit (ReLU) function is used as the non-linear projection function across different layers to prevent vanishing gradient and sparse activation problems. The sigmoid function is applied in the last layer. We use cross entropy loss and L2 regularization as the loss function and the Adam optimization to minimize the loss. We use Python Programming Language (Version3.8). Models are implemented with Python Pytorch[30] and HuggingFace Transformers.[31] The training was performed on an NVIDIA RTX A5000 (24GB RAM). *Our codes are available at https://github.com/weimin17/Multimodal_Transformer.*

Results

Prediction Results Analysis

We predict in-hospital mortality based on the first 48 hours of an ICU stay, which is a binary classification task. We use the same train-test setting defined in the benchmark[14] with 15% of the training data as a validation set, and similar

to Khadanga et al, we remove all clinical notes that do not have any chart time associated and patients that do not have any clinical notes. The statistics on the post-processed data are shown in Table 2.

Table 2. Statistics of the post-processed MIMIC III data for the in-hospital mortality prediction task.

	Train	Validation	Test
Negative	12216	2682	2748
Positive	1852	404	359
Total	14068	3086	3107

To comprehensively evaluate the performance of our model, we compute the metrics AUCROC, AUCPR, and F1. As the dataset is imbalanced, other metrics such as accuracy may be misleading. We run all experiments five times with different initialization and report the mean and standard deviation of the results.

Results in Table 3 demonstrate that our models outperform other methods in classifying in-hospital mortality. We achieve an AUCPR score of 0.538, an AUCROC score of 0.877, and an F1 score of 0.490.

Table 3. Experiment Results of different methods on MIMIC III In-Hospital Mortality Prediction Task.

	Prediction Model	AUCPR	AUCROC	F1
Only Variables	LSTM	0.460(+-0.013)	0.821(+-0.006)	0.392(+-0.038)
	Transformer	0.473(+-0.011)	0.827(+-0.005)	0.406(+-0.025)
Only Notes	MBERT	0.482(+-0.012)	0.851(+-0.005)	0.382(+-0.079)
Fusion	MBERT+LSTM	0.508(+-0.002)	0.859(+-0.001)	0.478(+-0.023)
	Multimodal Transformer (Ours)	**0.538(+-0.004)**	**0.877(+-0.001)**	**0.490(+-0.036)**

In the following section, we first investigate variants of BERT models with regard to pretraining and fine-tuning. Then, we visualize the important words in clinical notes by Integrated Gradient. Finally, we analyze the important clinical variables with the Shapley value.[32]

Domain adaptive pretraining and task adaptive fine-tuning on BERT models

In order to show the importance of domain adaptive pretraining and task adaptive fine-tuning in BERTs, we conduct an ablation study with only the pretrained models versus with fine-tuning using a single modality - clinical notes. The results are shown in Table 4. As expected, the general-purpose 'BERT' achieves the poorest result, whereas the MBERT achieves the best performance. These experiments suggest that clinical notes with proper trained language model are able to provide helpful information in clinical tasks, which enables deep learning techniques to leverage rich textual information to better understand the patient situation.

Table 4. Experiments on various Pre-trained and Fine-tuned BERTs. Use only MIMIC III clinical notes for in-hospital mortality prediction without considering clinical variables information. *Freeze* indicates only training the final classifier while keeping the BERT models unchanged. *Fine-tuned* indicates fine-tuning the BERTs for the in-hospital mortality downstream task.

	AUCPR	AUCROC	F1	AUCPR	AUCROC	F1
	Freeze			**Fine-tuned**		
BERT	0.182(+-0.016)	0.649(+-0.020)	0(+-0)	0.417(+-0.023)	0.829(+-0.005)	0.342(+-0.054)
BioRoBERTa	0.182(+-0.013)	0.661(+-0.016)	0(+-0)	0.455(+-0.010)	0.841(+-0.005)	**0.419(+-0.044)**
BioBERT	0.191(+-0.005)	0.664(+-0.011)	0(+-0)	0.444(+-0.027)	0.843(+-0.006)	0.377(+-0.045)
Clinical BERT	**0.265(+-0.006**)	**0.731(+-0.004**)	0(+-0)	**0.482(+-0.012)**	**0.851(+-0.005)**	0.382(+-0.079)

Clinical Notes Visualization and Interpretation

To provide an interpretation for the clinical notes and to better visualize the information, we evaluated the words that were important for prediction in our MBERT model using Integrated Gradients (IG).[34] We apply the IG method to study the problem of attributing the prediction of a deep network to its input features, as an attempt towards explaining

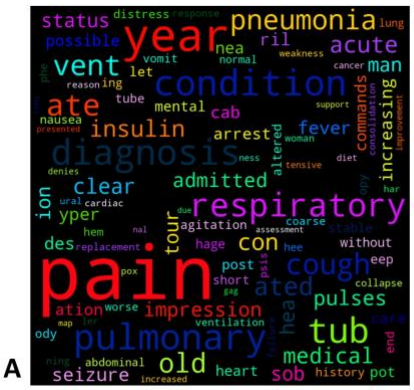
individual predictions. IG is computed based on the gradient of the prediction outputs considering the input words. Higher IG values indicate that a word is more important to the model's prediction, while smaller IG values indicate that a word is less important. We compute the IG value of all tokens in the clinical notes for all patients in the test data, and list the tokens with the highest IG values. Note that due to the BERT tokenization mechanism, the inputs are tokens instead of words. For example, the phrase "the patient has been extubated" would be tokenized to "the patient has been ex ##tub ##ated" as the input. To make the results more readable, we remove all the numbers, tokens that only have one or two characters, and separators in post processing. The tokens and their IG values are evaluated by a clinician for their clinical meaningfulness for mortality prediction. The tokens are sorted by those that are "Clinically Meaningful Indicators" of symptoms, prognosis, or care; "Unclear Tokens" which are difficult to attribute a single meaning to; and Headers/Common Words that are parts of structured notes or ubiquitously words used in medical notes, illustrated in Table 5.

Table 5. Top 20 Features Integrated Gradient Values Sorted for interpretability. The results come from patients in test data.

Rank	Clinically Meaningful Indicators	Unclear Tokens	Headers /Common Words	Rank	Clinically Meaningful Indictors	Unclear Tokens	Headers /Common Words
1	pain	##tub	with	11	sob	cab	possible
2	pulmonary	ate	year	12	status	##nea	let
3	respiratory	##ated	condition	13	seizure	##ation	without
4	vent	con	diagnosis	14	increasing	##pot	end
5	cough	##hea	old	15	fever	##eep	history
6	pneumonia	##tour	medical	16	arrest	##hage	from
7	insulin	##ion	impression	17	commands	##ody	not
8	acute	des	admitted	18	care	##ing	will
9	clear	##yper	the	19	heart	##opy	male
10	pulses	##ril	man	20	mental	hem	this

Several words with high IG values appear to be parts of structured headers, such as "medical condition," "diagnosis" or "impression," so are categorized separately from text that was unstructured. Additional words that are used ubiquitously in clinical notes, such as "year," "old," and "with" are also categorized separately as they were less likely to distinguish prognostic differences. Evaluating the top 20 clinically meaningful indicators that are important for mortality prediction, there are some interesting observations for clinical interpretation. "Pain", which is the indicator most important for prediction, is a common symptom in ICU care and can correlate with disease severity or disability. Indicators 2-6 correspond to pulmonary pathology, and the attribution of high importance to these indicators is in line with severity of respiratory illness and the need for ICU level care such as mechanical ventilation. Other indicators, such as "fever" or "seizure", are manifestations of acute illness, which could also have prognostic significance in predicting mortality. Clinical indicators such as "status," "commands," "mental," and "agitation" corresponded to mental status, and as delirium is associated with worse prognosis, it is not surprising that these indicators have prognostic importance in prediction.[35,36] Additional words such as "care" had multiple contexts when reviewing the notes; phrases such as "plan of care" or "resp care" are often used as headers, but used in other contexts it could be interpreted as a poor prognostic signal (e.g. "withdrawal of care") or a favorable prognostic signal (e.g. " response to care").

Figure 6.A is the word cloud visualization of the top 200 important words. We select the top 10 words with highest IG in every note, and compute all the notes. Says there are 10000 notes, then there would be 10*10000 top words (repeatable), and we compute the frequency of each unique word. The font size reflects the frequency. Figure 6.B is a demo illustration of word importance among the clinical notes.



B

True Label	Predicted Label	Word Importance
Death	Death	...impression : increase in cardiac failure with likely super ##im ##posed as ##piration pneumonia. ...
Death	Death	... she has a strong s ##pont non pro ##d cough and has not c / o any sob . g ##i - og ##t d ##c ' d when ex ##tub ##ated ...
Survival	Survival	... h ##ct has remained stable during night . re ##sp : remains oral ##ly in ##tub ##ated . becoming re ##sp al ##kal ##otic , therefore decreased im ##v to 10 with improvement of p ##h ...
Survival	Survival	..p ##t improved over next 45 minutes . c ##x ##r > ll in ##fi ##lt ##rate , no obvious pulmonary ed # #ema . as re ##sp distress improved ...

Figure 6. A: Word Cloud for clinical tokens with high IG values. Larger font indicates the word is more likely to appear as a top-ten IG value in clinical notes. B: Illustration of word importance in clinical notes based on IG value. The darker green color indicates the words that are more important (higher IG value) to the prediction, while the black color is background color.

Clinical Variables Feature Analysis

Next, we implement Shapley values to rank the important clinical variables. Shapley values[31] involve a game theory-based approach to explain the prediction of deep learning models. They measure the contribution of a given feature value to the difference from the actual prediction to the mean prediction. The top 10 out of 17 clinical variables (Table 6) show that for structured EHR data, the highest ranked variables also correlate with disease severity and poorer prognosis. These variables represent clinically important information such as mental status using the Glasgow Coma Scale, respiratory status and oxygenation, and hemodynamic measurements. They also provide interpretability of the directionality of impact for continuous variables, with poor prognostic variables like higher need for supplemental oxygen (Fraction inspired oxygen) increasing the likelihood for predicting death, and favorable prognostic variables, like higher blood systolic and blood pressure decreasing the likelihood of predicting death.

Table 6. Top 10 Features of Clinical Variables based on Shapley Value.

Rank	Shapley Value	Feature	Rank	Shapley Value	Feature
1	0.0374	Glascow Coma Scale total	6	0.0299	Diastolic blood pressure
2	0.0315	Fraction inspired oxygen	7	0.0296	Heart Rate
3	0.0312	Oxygen saturation	8	0.0288	Weight
4	0.0308	Glucose	9	0.0282	Mean blood pressure
5	0.0299	Glascow Coma Scale eye opening	10	0.0282	Systolic blood pressure

Discussion

Vast clinical datasets provide the opportunity for deep learning techniques to study the problem of in-hospital mortality prediction. Compared to previous related work, which mostly considers single modality or only naively concatenates embeddings from different modalities, our work demonstrates a novel way to integrate multimodal knowledge and leverage clinical notes information for better predictions. Meanwhile, the novel application of transformers on clinical data enables the model to consider information from all other time stamps when fusing the multimodal information because of the unique attention mechanism in the transformer block. To our best knowledge, this is the first work utilizing a transformer block to fuse clinical notes and clinical variable information while dealing with time series data in EHR data. We also conduct comprehensive experiments to demonstrate that our proposed method outperforms other methods by achieving high performance (AUCPR: 0.538, AUCROC: 0.877, F1:0.490).

The ablation study of domain adaptive pretraining and task adaptive fine-tuning on various BERTs verifies the significance of pretraining and fine-tuning when we implement the BERT models on natural language text, especially on domain-specific clinical notes.

The analysis and visualization of important words in clinical notes also provide interesting findings. The ranking of words by IG values provides face validity that many of the important words used for prediction are clinically related to diseases or processes that are prognostically important, such as severity of respiratory disease or mental status. Other words, such as "care," may be used in multiple contexts, and are more difficult to interpret as isolated words.

One important information source that could enhance our model's interpretability is considering the negation. The clinical meaning of notes can change significantly with negation, such as "crackles" indicating abnormal lung exam findings, and "not crackles" indicating a normal lung exam. In the future, we will employ techniques like the NegEx algorithm[37] to consider negation of key words to better explain the clinical words' meaning.

Conclusion

In this paper, we demonstrate a novel transformer based model, Multimodal Transformer, to leverage clinical notes and fuse multimodal knowledge from clinical data. We implement a transformer block to integrate both clinical notes and clinical variables while considering the time series information. The results demonstrate that our proposed Multimodal Transformer outperforms other methods. Additionally, we conduct different studies to further investigate the importance of domain adaptive pretraining and task adaptive fine-tuning for the Clinical BERTs. We also provide methods to interpret and visualize the important words in clinical notes using IG and Shapley methods, which demonstrate interesting findings on important features in clinical variables.

Acknowledgement

This research was partially supported by NSF grants IIS-1909038 and CCF-1855760.

References

1. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015. ONC data brief. 2016 May;35(35):2008-15.
2. Dalal AK, Piniella N, Fuller TE, Pong D, Pardo M, Bessa N, Yoon C, Lipsitz S, Schnipper JL. Evaluation of electronic health record-integrated digital health tools to engage hospitalized patients in discharge preparation. Journal of the American Medical Informatics Association. 2021 Mar 18;28(4):704-12.
3. Schwartz JM, Moy AJ, Rossetti SC, Elhadad N, Cato KD. Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review. Journal of the American Medical Informatics Association. 2021 Mar 1;28(3):653-63.
4. Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC medical informatics and decision making. 2020 Dec;20(1):1-0.
5. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. BMJ open. 2021 Jul 1;11(7):e044779.
6. Dong X, Rashidian S, Wang Y, Hajagos J, Zhao X, Rosenthal RN, Kong J, Saltz M, Saltz J, Wang F. Machine learning based opioid overdose prediction using electronic health records. InAMIA Annual Symposium Proceedings 2019 (Vol. 2019, p. 389). American Medical Informatics Association.
7. Yang B, Wu L. How to leverage the multimodal EHR data for better medical prediction?. InProceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 2021 Nov (pp. 4029-4038).
8. Cai X, Perez-Concha O, Coiera E, Martin-Sanchez F, Day R, Roffe D, Gallego B. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. Journal of the American Medical Informatics Association. 2016 May;23(3):553-61.
9. Teo K, Yong CW, Chuah JH, Hum YC, Tee YK, Xia K, Lai KW. Current Trends in Readmission Prediction: An Overview of Approaches. Arabian journal for science and engineering. 2021 Aug 16:1-8.
10. Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR medical informatics. 2017 May 9;5(2):e7235.
11. Sheikhalishahi S, Balaraman V, Osmani V. Benchmarking machine learning models on multi-centre eICU critical care dataset. Plos one. 2020 Jul 2;15(7):e0235424.
12. Rocheteau E, Tong C, Veličković P, Lane N, Liò P. Predicting Patient Outcomes with Graph Representation Learning. arXiv preprint arXiv:2101.03940. 2021 Jan 11.
13. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, Zheng WJ, Roberts K. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. Journal of Biomedical Informatics. 2021 Mar 1;115:103671.

14. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Scientific data. 2019 Jun 17;6(1):1-8.
15. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.
16. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
17. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964. 2020 Apr 23.
18. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.
19. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342. 2019 Apr 10.
20. Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: transformer for electronic health records. Scientific reports. 2020 Apr 28;10(1):1-2.
21. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23.
22. Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine. 2017 Nov 9;34(6):96-108.
23. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC medical informatics and decision making. 2020 Dec;20(1):1-1.
24. Khadanga S, Aggarwal K, Joty S, Srivastava J. Using Clinical Notes with Time Series Data for ICU Management. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 2019 Nov (pp. 6432-6437).
25. Deznabi I, Iyyer M, Fiterau M. Predicting in-hospital mortality by combining clinical notes with time-series data. InFindings of the Association for Computational Linguistics: ACL-IJCNLP 2021 2021 Aug (pp. 4026-4031).
26. Rahman W, Hasan MK, Lee S, Zadeh A, Mao C, Morency LP, Hoque E. Integrating multimodal information in large pretrained transformers. InProceedings of the conference. Association for Computational Linguistics. Meeting 2020 Jul (Vol. 2020, p. 2359). NIH Public Access.
27. Teixeira PL, Wei WQ, Cronin RM, Mo H, VanHouten JP, Carroll RJ, LaRose E, Bastarache LA, Rosenbloom ST, Edwards TL, Roden DM. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. Journal of the American Medical Informatics Association. 2017 Jan 1;24(1):162-71.
28. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine. 2020 Oct 16;3(1):1-9.
29. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3(1):1-9.
30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32.
31. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771. 2019 Oct 9.
32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
33. Lo K, Wang LL, Neumann M, Kinney R, Weld DS. S2ORC: The semantic scholar open research corpus. arXiv preprint arXiv:1911.02782. 2019 Nov 7.
34. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. International conference on machine learning 2017 Jul 17 (pp. 3319-3328). PMLR.
35. Girard TD, Pandharipande PP, Ely E. Delirium in the intensive care unit. Critical care. 2008 May;12(3):1-9.
36. Salluh JI, Wang H, Schneider EB, Nagaraja N, Yenokyan G, Damluji A, Serafim RB, Stevens RD. Outcome of delirium in critically ill patients: systematic review and meta-analysis. bmj. 2015 Jun 3;350.
37. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001 Oct 1;34(5):301-10.