Predicting in-hospital mortality by combining clinical notes with time-series data

Iman Deznabi, Mohit Iyyer, Madalina Fiterau
College of Information and Computer Sciences
University of Massachusetts Amherst
{iman, miyyer, mfiterau}@cs.umass.edu

Abstract

In intensive care units (ICUs), patient health is monitored through (1) continuous vital signals from various medical devices, and (2) clinical notes consisting of opinions and summaries from doctors which are recorded in electronic health records (EHR). It is difficult to jointly model these two sources of information because clinical notes, unlike vital signals, are collected at irregular intervals and their contents are relatively unstructured. In this paper, we present a model that combines both sources of information about ICU patients to make accurate in-hospital mortality predictions. We apply a fine-tuned BERT model to each of the patient's clinical notes. The resulting embeddings are then combined to obtain the overall embedding for the entire text part of the data. This is then combined with the output of an LSTM model that encodes patients' vital signals. Our model improves upon the state of the art for mortality prediction, attaining an AUC score of 0.9, compared to the previous 0.87, setting a new standard for mortality prediction on the MIMIC III benchmark.[1]

1 Introduction

One of the major costs in healthcare is critical care in intensive care units. A crucial aspect of critical care is mortality prevention. With advancements in healthcare, a patient's condition during an ICU visit is monitored through devices that measure many different vital signals, such as heart rate, systolic blood pressure, temperature, etc. Additionally, we have access to clinical notes written by medical professionals during the patient's stay. This data can be used to predict the condition of patients during their ICU stay, which can help in managing the expensive resources in hospitals and providing these services to patients who need them most in a more cost-effective way.

Most prior work has been focused on predicting patient health using the time-series data gathered from medical devices (Lipton et al., 2016; Che et al., 2018; Narayan Shukla and Marlin, 2020; Xu et al., 2018). More recent work also attempted to leverage clinical notes for making these predictions (Zhang et al., 2019; Ghorbani et al., 2020; Lee et al., 2020; Alsentzer et al., 2019). However, research on combining time series data with clinical notes for outcome prediction has been limited due to the irregularity of the clinical notes compared to the time series data, and the complexity of jointly modeling the two sources of information.

In this work, we propose a multimodal neural network that combines time-series data from medical devices with textual information from clinical notes to improve the prediction of in-hospital mortality. This task is defined as predicting whether a patient will die before getting discharged from hospital based on the first two days of data. Our model improves on the prior state of the art by using fine-tuned BERT-based models (Devlin et al., 2019) to encode clinical notes and integrates the resulting representations with time-series embeddings derived from a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997). We also show that fine-tuning clinical BERT models in isolation from other parts of the model and specifically for in-hospital prediction tasks brings additional performance improvements. Furthermore, separately modeling each clinical note and combining the resulting embeddings leads to our final model which supersedes the state-of-the-art in predicting "in-hospital mortality" in terms of AUC score.

[1] Our implementation is available at https://github.com/Information-Fusion-Lab-Umass/ClinicalNotes_TimeSeries

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4026–4031
August 1–6, 2021. ©2021 Association for Computational Linguistics

2 Related work

Most of the previous work in mortality prediction in critical care is focused on utilizing patient vital signals as time-series data for making predictions. Harutyunyan et al. (2019) provided a benchmark and defined 4 tasks based on MIMIC III dataset (Johnson et al., 2016) for comparing these models. This paper used a recurrent neural network to obtain the predictions. Other models also use recurrent neural networks in different forms to predict outcomes in healthcare data (Liu and Chen, 2019; Suresh et al., 2018; Lipton et al., 2016). Furthermore, others (Che et al., 2018; Shukla and Marlin, 2019; Narayan Shukla and Marlin, 2020; Horn et al., 2020) used the irregular nature of the data over time in their models.

For clinical notes, some models used pretrained embedding models (Zhang et al., 2019; Chen et al., 2019), while some use various other machine learning techniques for outcome prediction on data collected from the ICU (Jin et al., 2018; Boag et al., 2018; Ghorbani et al., 2020).

More recently, BERT-based models have been adopted for this domain following their incredible success in Natural Language Processing (NLP). A number of studies trained and used BERT-based models for clinical applications (Lee et al., 2020; Huang et al., 2019; Darabi et al., 2020; Li et al., 2020; Alsentzer et al., 2019).

All these models only use one source of available data when predicting medical outcomes. However, Khadanga et al. (2019) showed the usefulness of combining time-series and clinical notes for ICU outcome prediction. They used a convolution neural network (CNN) on top of pretrained word embedding from (Zhang et al., 2019) for getting a representation of clinical notes and a long short-term memory network (LSTM) for embedding the time series part of the data. The two representations were then concatenated to make the predictions. Concurrently to our work, Yang et al. (2021) also showed the usefulness of combining time-series data with information from clinical notes. They used an LSTM model for the time-series part of the data and use a convolutional neural network with label-aware attention layer for the clinical notes. We will use the same intuition for combining the representation of clinical notes and the time-series part of the patients' data to improve the performance of in-hospital mortality prediction. In the next section we define the notations and our method architecture.

3 Method

In this section, we first introduce the notation for our task and then describe our proposed model. Each patient has several clinical notes collected during their stay. For a patient p with N clinical notes, we denote $C^{(p)} = \{c_{t_1}^{(p)}, c_{t_2}^{(p)}, ..., c_{t_N}^{(p)}\}$ for clinical notes collected at times $\{t_1, t_2, ..., t_N\}$. We also have time-series data collected during the patient's stay as $X_{1:T}^{(p)} = \{x_1^{(p)}, x_2^{(p)}, x_3^{(p)}, ..., x_T^{(p)}\}$ where $x_i^{(p)}$ is the data collected at index i. For simplicity, the patient index p will be dropped for the rest of the section.

Our model brings three major contributions compared to prior work in this area. First, we use a fine-tuned BERT-based model for modeling clinical notes as these new attention-based models trained on language models such as BERT have been shown to significantly outperform previous approaches in many NLP applications (Devlin et al., 2019) including clinical note prediction (Lee et al., 2020; Alsentzer et al., 2019; Huang et al., 2019). Second, we feed each clinical note of a patient separately to the text part of our model, and then combine the resulting embeddings to get the final representation of text part of the data. We argue that since the clinical notes are collected separately, and usually from different sources, this approach is preferable to concatenating the text of the clinical notes for each patient and then feeding this to the model. Finally, we also fine-tune the BERT-based models in isolation on "in-hospital mortality" task, which brings additional performance improvements.

The model architecture is shown in Figure 1. It consists of two parts. The first part obtains the embedding of clinical notes and combines them to form a final representation of all the clinical notes, for which we use the notation H_C. The other part of the model is responsible for encoding the time-series part of the data, which will be referred to as H_X. These representations will then be concatenated to form the final representation of the patient $H = H_C \oplus H_X$ where $\oplus$ is the concatenation operation. Finally, a softmax layer is applied on top of H to get the final predictions.

3.1 Clinical Notes Representations

Modeling clinical notes require capturing interactions between distant words (Huang et al., 2019)

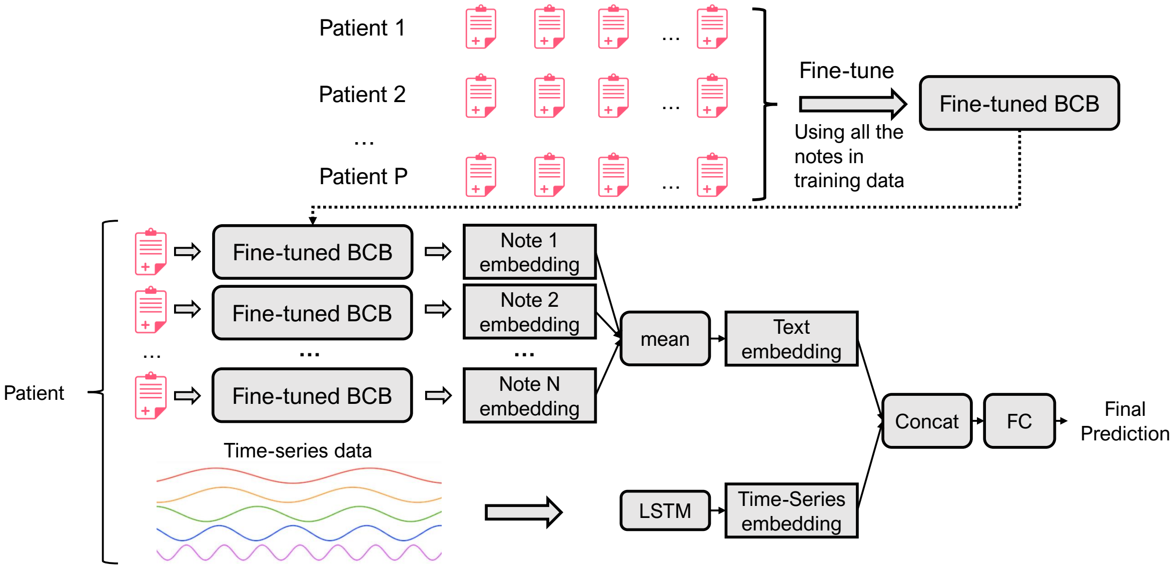

Figure 1: The model architecture. First, the Bio+Clinical BERT (BCB) (Alsentzer et al., 2019) model is fine-tuned for the in-hospital mortality prediction task using all the available notes in the training data. Then the average embedding of patient notes are combined with the time-series embedding to obtain the final patient representation. Then a fully connected layer is used to predict patient's in-hospital mortality.

and BERT-based models are designed to capture these interactions. Furthermore, Huang et al. (2019); Lee et al. (2020); Alsentzer et al. (2019) showed that these models outperform the traditional NLP models in this area. Considering this success we use a BERT-based model fine-tuned to clinical notes for getting the embedding of each clinical note of the patient. Specifically, we used the Bio+Clinical BERT model described in Alsentzer et al. (2019) which fine-tuned Bio BERT(Lee et al., 2020) on all the clinical notes in MIMIC III dataset.

After initializing our text model to Bio+Clinical BERT, we added a softmax layer on top of the classification token (CLS) output of each clinical note separately and the model is fine-tuned further on the task of predicting the in-hospital mortality of the patient. Finally, to obtain the overall text embedding of a patient, we take the average embedding of all the clinical notes available for the patient. This is shown in equation 1.

$$H_C = \frac{1}{N}\sum_{i=1}^{N} \mathrm{BCB}(c_{t_i}) \qquad (1)$$

where the $\mathrm{BCB}(c_{t_i})$ function is the output of the CLS token of the fine-tuned Bio+Clinical BERT model when c_{t_i} is provided as an input to the model and H_C is the aggregated representation of all the clinical notes of the patient.

3.2 Time series

For the time-series part of the model, following the works by Harutyunyan et al. (2019) and Khadanga et al. (2019), we first limit the time-series data to first 48 hours of the patient stay, then we resample the time-series data to intervals of 1 hour. If there are any duplicate values in that hour for a variable, we use the most recent values. We use forward imputation for missing values. If no previous value is recorded we use the pre-set values for the features given in Harutyunyan et al. (2019). After pre-processing the time series data, we used an LSTM network. We input the whole time-series data of a patient ($X_{1:T}$) to the LSTM model and use the final hidden state as the representation of time-series part of the patient data.

$$H_X = \mathrm{LSTM}(X_{1:T}) \qquad (2)$$

4 Results

In this section, we will show the results achieved on the in-hospital mortality prediction task on MIMIC III dataset. To ensure consistency with previous work and the benchmarks presented on Harutyunyan et al. (2019), we used the same preprocessing steps as defined by Harutyunyan et al. (2019) and split the patients into train, validation and test sets using the same splits. We also followed the preprocessing steps of Khadanga et al. (2019) to process the clinical notes and removed, from the dataset,

Type	Model	AUC
Only time-series	LSTM*	0.835
Only text	CNN*	0.831
	BERT	0.671
	BCB	0.760
	BCB FT	0.835
	BCB FT MN	**0.875**
Text + Time-series	CNN + LSTM*	0.867
	BERT + LSTM	0.840
	BCB + LSTM	0.851
	BCB FT + LSTM	0.873
	BCB FT MN + LSTM	**0.899**

Table 1: Area under the receiver operating characteristic (AUC) results on the held out data for all models including baselines. The results are averages of 5 different initializations of the models. The models indicated with a star (*) are using the same architecture as described in Khadanga et al. (2019). BCB is short for Bio+Clinical BERT (Alsentzer et al., 2019), FT means fine tuned separately for in-hospital mortality task. The baseline models only use the first available clinical note while models indicated with MN (multi-note) use all the available clinical notes in first 48 hours of a patient's ICU stay.

patients who do not have any associated medical notes. We also only used the clinical notes collected at first 48 hours of patients' stay. After these steps, our dataset consisted of 11579 records in the training set, 2570 in the validation set, and 2573 in the test set.

The time-series part of the model uses all 17 medical variables (e.g., heart rate, height, and Glucose) recorded during the first 48 hours of the patients' stay. Since in-hospital mortality prediction is a binary classification task with unbalanced classes (only around 10% of the patients suffered mortality in this dataset), the area under the receiver operating characteristic (AUC) is used for evaluating our models. The best performing model uses one layer of LSTM cells with 256 units, an Adam optimizer (Kingma and Ba, 2015) with initial learning rate of 2×10^{-5} for training and a weight decay of 0.01. The models are implemented with the Pytorch library (Paszke et al., 2019) and we run all the experiments 5 times with different initialization and report the mean of the results.

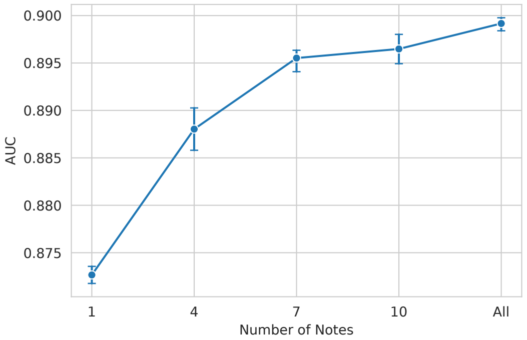

Figure 2: Increase in full model performance as more clinical notes are included in order of time for each patient. The Y axis shows the AUC score of the model on the test data and the X axis shows the number of included clinical notes. Each bar indicates 95% confidence interval over 5 different runs of the model.

Table 1 summarizes the results achieved by our models and the baseline models that use only time-series data, use a simple BERT instead of the Bio+Clinical BERT model, and the results with and without fine tuning Bio+Clinical BERT model further for the in-hospital mortality prediction task. The models which use both time-series and text data are also shown. Moreover, the results are compared with Khadanga et al. (2019) which, to the best of our knowledge, is the state-of-the-art for in-hospital prediction task in MIMIC III benchmarks. Our final model, which uses fine-tuned clinical BERT model for the text part of the data and an LSTM model for time-series part (shown in Figure 1) significantly outperforms the baseline models as well as the state-of-the-art models.

To assess the value of using multiple notes for each patient, we experimented by including various numbers of clinical notes for each patient in our final model. The AUC results are shown in Figure 2. It is apparent that the model's performance improves with more notes, and the best performance is achieved when all the notes are used.

Although we lose some information contained in the clinical notes by truncating them to fit into the maximum acceptable length of the model, using the entire text of the clinical notes by segmenting them into chunks and taking the average did not yield significant improvements in the performance of our final model. In our experiments, providing all chunks either by segmenting the concatenated text of the clinical notes or individual notes separately only improved the final AUC score from 0.899 to a maximum of 0.902 while taking approximately twice as long to train.

5 Conclusions and Future Work

Improved prediction of the admission outcome in Intensive Care Units (ICUs) can be tremendously valuable in supporting decisions in clinical diagnosis. Previous work mostly focused on using the patients' vital sign information recorded during their stay as time-series data for these predictions. However, the clinical notes written by healthcare workers include important information on both the history and current conditions of the patients in the hospital, which can be leveraged to improve these predictions significantly. In this work, we proposed a novel method for leveraging the information available both in clinical notes and time-series data. Our method consists of fine-tuning Bio+Clinical BERT model for in-hospital mortality prediction task and then combining it with the available time-series data. Finally, we showed that using all the clinical notes available significantly enhances the ability to make these kinds of predictions.

In the future, we plan to improve the performance of the model by improving the time-series part of the model from an LSTM network to a model that can leverage irregular time-series data. Furthermore, the same model architecture can be used for other tasks discussed in the literature for such medical datasets such as physiologic decompensation or forecasting length of stay.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

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.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Sajad Darabi, Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. 2020. Taper: Time-aware patient ehr representation. *IEEE Journal of Biomedical and Health Informatics*.

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.

Ramin Ghorbani, Rouzbeh Ghousi, Ahmad Makui, and Alireza Atashi. 2020. A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. *IEEE Access*.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96.

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

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. 2020. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR.

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

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

A Johnson, T Pollard, and R Mark III. 2016. The mimic-iii clinical database. *PhysioNet. doi*, 10:C2XW26.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, Hong Kong, China. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and

Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific Reports*, 10(1):1–12.

Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. 2016. Learning to diagnose with LSTM recurrent neural networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Jiexi Liu and Songcan Chen. 2019. Non-stationary Multivariate Time Series Prediction with Selective Recurrent Neural Networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11672 LNAI, pages 636–649.

Satya Narayan Shukla and Benjamin M Marlin. 2020. Multi-Time Attention Networks for Irregularly Sampled Time Series. Technical report.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Satya Narayan Shukla and Benjamin M. Marlin. 2019. Interpolation-Prediction Networks for Irregularly Sampled Time Series. *7th International Conference on Learning Representations, ICLR 2019*.

Harini Suresh, Jen J Gong, and John V Guttag. 2018. 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*, pages 802–810.

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.

Haiyang Yang, Li Kuang, and FengQiang Xia. 2021. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):1–14.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.