Artificial Intelligence In Medicine 117 (2021) 102112

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Improving clinical outcome predictions using convolution over medical entities with multimodal learning

Batuhan Bardak, Mehmet Tan *

Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey

ARTICLE INFO

Keywords:
Deep learning
Healthcare
EHR
NER
Multimodal

ABSTRACT

Early prediction of mortality and length of stay (LOS) of a patient is vital for saving a patient's life and management of hospital resources. Availability of Electronic Health Records (EHR) makes a huge impact on the healthcare domain and there are several works on predicting clinical problems. However, many studies did not benefit from the clinical notes because of the sparse, and high dimensional nature. In this work, we extract medical entities from clinical notes and use them as additional features besides time-series features to improve proposed model predictions. The proposed convolution based multimodal architecture, which not only learns effectively combining medical entities and time-series Intensive Care Unit (ICU) signals of patients but also allows to compare the effect of different embedding techniques such as Word2vec and FastText on medical entities. Results show that the proposed deep multimodal method outperforms all other baseline models including multimodal architectures and improves the mortality prediction performance for Area Under the Receiver Operating Characteristics (AUROC) and Area Under Precision-Recall Curve (AUPRC) by around 3%. For LOS predictions, there is an improvement of around 2.5% over the time-series baseline. The code for the proposed method is available at https://github.com/tanlab/ConvolutionMedicalNer.

1. Introduction

Electronic Health Record (EHR) data collected from patients who have been admitted into hospitals or Intensive Care Units (ICU) offer a detailed overview of patients consisting of but not limited to demographics, insurance, laboratory test results, and medical notes. With the EHR data becoming available for researchers, there has been increasing interest in using it with deep learning algorithms. Besides rapid progress in deep learning area, after Medical Information Mart for Intensive Care (MIMIC-III) [1], today's most popular public EHR database, was released, numerous studies have achieved successful results using this data set in predicting different clinical outcomes [2–4].

Understanding the health condition of the patient by observing the clinical measurements, laboratory test results and predicting the condition of patients during their ICU stay is a vital problem. In this paper, we focus on two different common risk prediction tasks, mortality (in-hospital & in-ICU) and length of ICU stay (LOS). Both are very important clinical outcomes for determining treatment methods, planning hospital resources, and ultimately saving lives. Previous studies primarily focused on predicting clinical events using only the structured data of patient such as historical patient diagnosis (ICD codes) [5,6], lab results and patient ICU measurements [7–9] and did not benefit from the unstructured data in EHR. The EHR data which consists of clinical notes written by doctors, nurses, or radiologists, discharge notes, and many other sources, contains quite detailed information about patients, projecting the knowledge and inference of doctors and even critical details about patient health status for many cases. As per the importance of the clinical notes, researchers want to take advantage of the rich content in clinical notes. Moreover, with the recent developments in Natural Language Processing (NLP), there has been an increasing interest in using clinical notes to make clinical model predictions [10,11]. Although it may be possible to leverage clinical notes to make more accurate predictions, these notes may consist of long written free-text with an unusual grammatical structure and may contain redundant information. As it may be hard to process raw clinical notes, because of their high-dimensional and sparse nature, extracting medical entities is required to unlock the medical information trapped in the clinical notes and to feed them into prediction models.

Named Entity Recognition (NER) is a fundamental task in NLP that focuses on information extraction aiming to extract entities in a text and

* Corresponding author.
E-mail address: mtan@etu.edu.tr (M. Tan).

https://doi.org/10.1016/j.artmed.2021.102112
Received 10 July 2020; Received in revised form 18 April 2021; Accepted 11 May 2021
Available online 13 May 2021

classify them into predefined classes. These classes can be locations, people, or organizations in general NER algorithms [12,13]. There can be various NER models for different domains like cybersecurity [14] or medicine [15]. Recently, several deep learning algorithms were applied to clinical texts to train clinical Named Entity Recognition models. These clinical NER models generally try to extract medical information such as disease, drugs, dosage and frequency.

In this paper, we argue that the integration of structured data in EHR and medical entities positively affects the prediction of mortality and LOS. Also, the effect of different word representations such as Word2Vec [16], FastText [17] and the concatenation of both on medical entities are investigated. To evaluate the success of the proposed multimodal architecture, first models are trained separately with structured and medical entity features. Then we apply multimodal approach and use these features together in several ways to show the effectiveness of the proposed network. The results indicate a promising increase in performance on mortality and LOS tasks when the medical entities are used with structured data in a multimodal approach. To sum up, the main contributions of this work can be listed as follows.

- **Comparing clinical word embeddings.** Representing the medical entities is a critical problem and there are various word embedding methods that capture different semantic and syntactic features about the same word. In this study, different types of word embedding methods (Word2Vec, FastText, Concatenation) are experimented and the outcomes of these methods on clinical tasks are discussed.
- **Embedding techniques for medical entities.** To find an efficient way for embedding medical entities, we make experiments with different methods such as average representation, Doc2Vec, and 1D Convolutional Neural Networks (CNN). The experimental results show that convolutional based method is the best way to embed medical entities among the candidate methods.
- **Novel pipeline for mortality and LOS problems.** We work with four different clinical outcomes such as in-hospital mortality, in-ICU mortality, LOS >3 and, LOS >7. To make successful predictions on these clinical tasks, we propose a novel, fully reproducible, and convolutional based deep multimodal pipeline.

In the next section, similar studies that work on clinical domain especially predicting mortality and length of stay at ICU are summarized. Following that, the data set, problem definitions, and deep learning models used in this study are discussed. Then, the experimental results are reported and the paper is concluded in the last section.

2. Related work

With the rapid development of deep learning algorithms in the last decade, the number of deep learning models increased substantially for various clinical predictions. Several studies have explored EHRs to solve clinical problems, e.g., Lipton et al. [18] used 13 different vital measurements to classify 128 diagnoses using Long Short Term Memory (LSTM) and DoctorAI [5] used Gated Recurrent Unit (GRU) to predict multi-label diagnosis for the next visit. Choi et al. [19] proposed early heart failure detection using Recurrent Neural Networks (RNNs). Forecasting the LOS and mortality have been a popular clinical problem for healthcare researchers in recent years. In earlier studies [20–22] on mortality prediction, hand-crafted features are selected and used simple machine learning models like logistic regression with different severity scores such as APACHE [23], SAPS-II [24], and SOFA [25]. Nowadays with the progress on deep learning, different architectures have been applied on EHR data to predict these kind of problems. Awad et al. [26] used ensemble learning to make an early mortality prediction and Sadeghi et al. [27] proposed a method to predict mortality using 12 features extracted from the vital signals in the first hour of ICU admission. Darabi et al. [28] used convolutional neural network to predict long-term mortality risk on the MIMIC-III dataset. More recent work [8] includes attention to their deep learning model to improve models' success. Another work [29] try to predict LOS for acute coronary syndrome patients. There is a comprehensive survey on mortality prediction and LOS [30]. Despite these studies and developments, one of the major problems that the healthcare researchers experienced, the researches on the literature are short of standardized preprocessing steps such as unit conversion, handling outlier and missing values, and transforming raw structured data into usable hourly time series data. In order to solve this problem, these studies [31–33] carried out a comprehensive benchmark on MIMIC-III for various tasks such as mortality, LOS, readmission, phenotyping and make their code publicly available. Purushotham et al. [33] extracts 17 features from the MIMIC-III and works on hospital mortality, LOS and ICD-9 code group predictions. They compared their proposed super learner method with feedforward and recurrent neural network. Another research [31] benchmarked their results on the MIMIC-III. They used multi-task learning approaches to predict four clinical prediction tasks such as risk of mortality, LOS, detecting physiologic decline, and phenotype classification. MIMIC-Extract [32] is the most recent work which is an open source pipeline for transforming MIMIC-III data into directly usable features. Their pipeline first transforms the raw vital sign and laboratory data into hourly time series and then apply some preprocessing steps such as unit conversion, outlier handling and imputing missing data. In this study, to increase reproducibility, MIMIC-Extract pipeline is used to featurize MIMIC-III data.

Medical entities which are extracted from clinical notes are used to improve proposed model predictions. Clinical natural language processing and information extraction has been widely studied in recent years on clinical notes. Two studies [34,35] proposed a deep learning based multi-task learning to make clinical predictions from clinical notes. Boag et al. [11] compared different embedding approaches such as Bag of Words (BoW), Word2Vec and LSTM on clinical note representation by evaluating the prediction performance on diagnosis prediction and mortality risk estimation. More recently, transformer-based architectures such as BERT [36], XLNET [37] gave state-of-the-art performance on different NLP tasks. These models are pre-trained on medical data, which is then fine-tuned on clinical text [38,39]. However, clinicians generally use medical jargon and shorthands when they take these clinical notes which makes it hard to process directly. There are a number of studies in the field of clinical NLP which try to extract medical entities in clinical notes [40–42]. In this work, we use med7 [15] which is developed for free-text Electronic Health Record. Then, these medical entities are combined with structured data to benefit from multimodal approach. For a detailed overview on deep learning for natural language processing in the clinical domain, readers can refer to [43].

Multimodal learning is a key research area that uses multiple sources to predict unique tasks [44]. This approach has shown success in image captioning tasks [45], visual question answering [46] and speech recognition [47]. In the healthcare research domain, Khadanga et al. [48] combines unstructured clinical notes and structural time-series data for predicting in-hospital mortality, decompensation, and LOS. Similarly, Shukla and Marlin [49] made unified mortality prediction and try to explore how physiological time series data and clinical notes can be integrated. The study by Jin et al. [50] is the closest to this work in terms of motivation. They made hospital mortality prediction by combining clinical notes and time series data. Clinical notes are represented with Doc2VecC [51] algorithm in two different ways. First, they directly combine clinical notes with time series data, second, they use neural network based clinical NER service to extract five types of medical entities and identify negated entities from clinical notes. After this pre-processing, they use the same representation with the first model and reported a 2% increase in the Area Under their Curve (AUC).

3. Materials and methods

In this section, we begin by describing the dataset. Next, the details of baseline models and clinical NER model are discussed. Finally, the

proposed multimodal deep learning models are explained.

3.1. Data

All models are trained on a publicly available MIMIC-III dataset which contains de-identified EHR data of 58,976 unique hospital admissions, 61,532 ICU admissions from 46,520 patients in the ICU of the Beth Isreal Deaconess Medical Center between 2001 and 2012. MIMIC-Extract [32], an open source data extraction pipeline, is used to extract structured time series features in MIMIC-III. MIMIC-Extract mainly focuses on the patient's first ICU visit with some patient inclusion criteria. They eliminate data from patients younger than 15 years old and where the LOS is not between 12 h and 10 days. This pipeline produces a cohort of 34,472 patients and 104 clinically aggregated time-series variables. In all experiments, we use the first 24 h of patient's data after ICU admission and only consider the patients with at least 30 h of present data like MIMIC-Extract. In the proposed multimodal approach, medical entities and time-series features are combined to be used together. Before applying the clinical NER model on notes, discharge summaries are dropped to avoid any information leak. Furthermore, clinical notes without the chart time information are eliminated. After these steps, we drop all patients who do not have any clinical notes in 24 h. The preprocessing on clinical notes are performed similar to [48]. In the train-test split, for all clinical tasks, we split the data based on class distribution with 70%/10%/20% ratio. Statistics of the final cohort and the others are summarized in Table 1.

Problem Definition. We mainly focus on two vital clinical prediction tasks, mortality(in-hospital & in-ICU) and LOS (>3 & >7) at ICU. The same definitions of the benchmark tasks defined by MIMIC-Extract are used as the following four binary classification tasks. The explanation of these tasks and the class distributions are as follows:

1. **In-hospital mortality**: Patient who dies during hospital stay after ICU admission (Significantly imbalanced, %10.5).
2. **In-ICU mortality**: Patient who dies during ICU stay after ICU admission (Significantly imbalanced, %7).
3. **Length-of-stay >3**: Patient who stays in the ICU longer than 3 days (Slight imbalanced, %43.2).
4. **Length-of-stay >7**: Patient who stays in the ICU longer than 7 days (Significantly imbalanced, %7.9).

3.2. Baseline models

In this subsection, the time-series baseline model that is evaluated on each of four benchmark tasks is discussed. Furthermore, the clinical NER model, embedding approaches to represent medical entities, and the multimodal baselines used in this study are explained.

3.2.1. Time series model

We employ both Long Short Term Memory (LSTM) [52] and Gated Recurrent Units (GRU) [53] networks to capture the temporal information between the patient features. As a result of time-series baseline experiments, GRU has shown a better AUC and AUPRC performance than LSTM up to %0.5–%1, while using a simpler architecture. Therefore, GRU is used for all of the multimodal architectures. In general, GRU cell has two gates, a reset gate r and an update gate z. With these gates, GRU can handle the vanishing gradient problem.

The mathematical formulation of GRU model can be iterated as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$\widehat{h}_t = tanh(W_h x_t + r_t \circ U_h h_{i-t} + b_h)$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \widehat{h}_t$$

$$\widehat{\text{prediction}} = \text{sigmoid}(W_h h_t + b_h)$$

where z_t and r_t respectively represent the update gate and the reset gate, $\widehat{h}_t$ the candidate activation unit, h_t the current activation, and $\circ$ represents element-wise multiplication. For predicting the mortality and LOS, a sigmoid classifier is stacked on top of the one layer GRU with 256 hidden units.

3.2.2. Multimodal approaches

In this work, besides time series features, information from clinical notes is used to improve clinical task prediction performance. Instead of working directly with clinical notes, we first aim to extract medical related keywords. Recently, there are some notable works in the clinical domain that made their pre-trained clinical NER models publicly available [54,55,15]. We use a pre-trained clinical NER model, med7 [15], which uses the same dataset that we use in experiments, MIMIC-III. This clinical NER model extracts seven different named entities such as 'Drug', 'Strength', 'Duration', 'Route', 'Form', 'Dosage', 'Frequency'. To represent the patient's medical entities, we try two different embedding methods, word embedding and document embedding. First, three different word embedding algorithms are used to represent each clinical NER model outputs and compare their performance. Second, Doc2Vec [56] algorithm is trained to represent the whole documents consisting of medical entities. The detailed schema of these two approaches are shown in Fig. 1 and the statistics of the extracted medical entities by med7 in MIMIC-III dataset for selected patients are shown in Table 2.

Word Embeddings. Different word embedding methods might capture various semantic features on the same word. In the experiments, to understand this variety, the performance of Word2Vec, FastText, and the concatenation of Word2Vec & FastText embeddings are compared. Word2Vec [16] is a two-layer neural network that learns the representations of words in the given text with two ways: as a continuous bag-of-words (CBOW) and as a skip-gram. FastText [17] is an extension of the skip-gram model implemented by Facebook's AI Research (FAIR) lab which can handle out-of-vocabulary (OOV) words, and can learn better representations for rare words using several n-grams for words. We use pre-trained Word2vec ($w_i \in \mathbb{R}^{100}$) and FastText embeddings ($f_i \in \mathbb{R}^{100}$) which was trained on 2.8 billion words from MIMIC-III clinical notes as shown in [38]. Lastly, an experimental embedding approach is designed which concatenates the Word2Vec and FastText representations horizontally ($c_i \in \mathbb{R}^{200}$). When the Word2Vec embedding does not exist for a given word, we make zero padding in this setting.

Document Embeddings. Doc2Vec is an extension of Word2Vec model to learn document-level embeddings instead of word level. Before learning document level representations, the first 24 h of patient's clinical notes are combined and clinical NER algorithm is applied on them to keep only medical related keywords in the clinical notes. When training Doc2Vec, the context window size is selected of 5 words. This algorithm produces the fixed-length feature vector ($d_i \in \mathbb{R}^{100}$) for each patient.

Table 1
Summary statistics of the original MIMIC-III dataset, and the final cohort that is used in this study.

	# of Patient	# of hospital admission	# of ICU admission
MIMIC-III (>15 years old)	38,597	49,785	53,423
MIMIC-Extract	34,472	34,472	34,472
MIMIC-Extract (at least 24 + 6 (gap) hours patient)	23,937	23,937	23,937
Final cohort (After clinical note elimination)	21,080	21,080	21,080

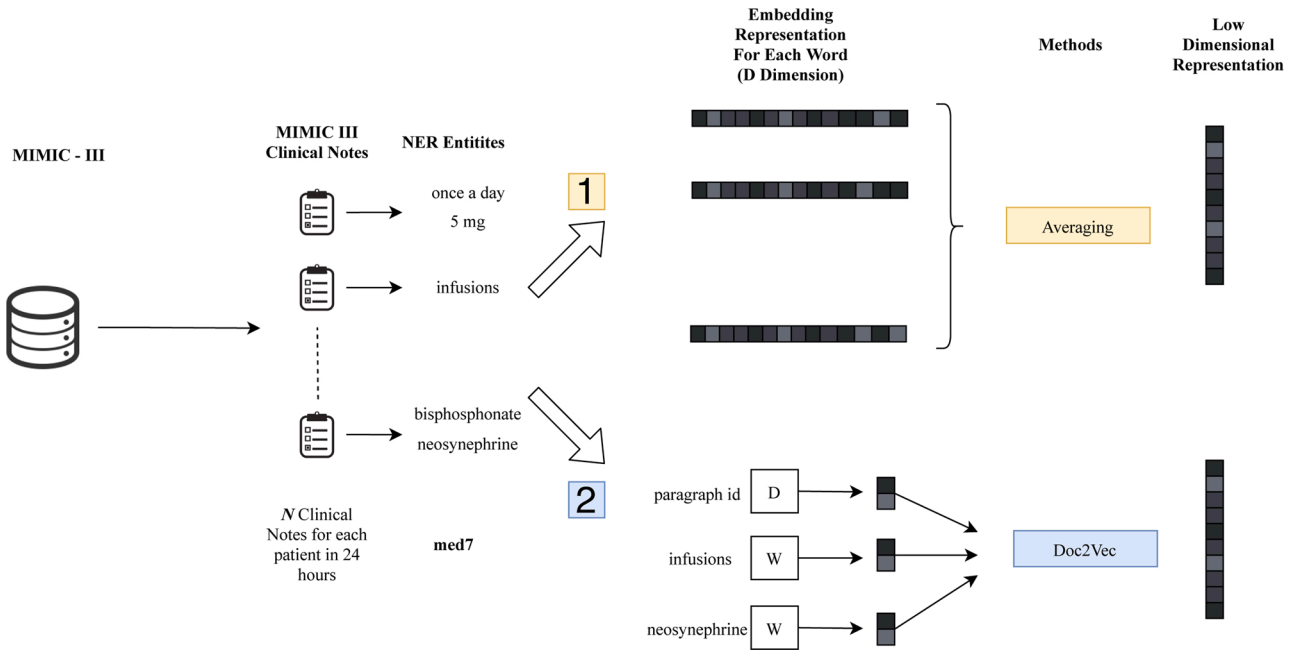

Fig. 1. Methodology for learning medical entity vectors. (1) The medical entities that are extracted from clinical notes are embedded into continuous word vectors. Then, we take the mean of these learned entity representations. (2) The words are removed from clinical notes if they are not belong to any medical entity category. Then, Doc2Vec is trained on the preprocessed clinical notes to learn low dimensional representation of medical entities.

Table 2
The first column shows the type of medical entity, the second columns shows the total number of related entity found in clinical notes, and the third column shows the number of unique entity number. The last column shows the output of med7 for example sentence given from clinical notes.

Medical entity	Total count	Unique count	Example
Drug	744,778	18,268	Magnesium
Strength	156,486	10,749	400 mg/5 ml
Form	40,885	597	suspension
Route	207,876	1193	PO
Dosage	126,756	7239	30 ml
Frequency	71,285	3344	bid
Duration	5939	1185	next 5 days

In this study, two different baseline multimodal approaches are presented with word and document embeddings that combine time-series data and medical entities.

Multimodal with Average Representation. This model takes the average of all medical entities associated with a patient. For each patient, there are N clinical notes and K medical entities extracted from these N clinical notes. Each medical entity is represented by a word embedding which is explained in Word Embeddings section. We sum K n-dimensional clinical entity representations component-wise and then divide this by K. Two different input types are used to train multimodal model. Time series data is processed through one layer GRU layer with 256 hidden units as explained in Section 3.2.1. Averaged representations of medical entities are combined with time-series feature maps that are learned via GRU. In the end, these merged feature representations are fed into the fully connected layer with 256 neurons, and a sigmoid classifier is added to the model.

Multimodal with Doc2Vec Representation. In this multimodal approach, instead of averaging medical entities, Doc2Vec algorithm is trained to obtain the fixed-length feature vector. First, we concatenate N clinical notes for each patient and discard keywords from these notes if the keyword is not a medical entity. Then the Doc2Vec algorithm is applied to learn a low level representation from notes for each patient. After the learning fixed-length feature vector, the same architecture as the average embedding approach is used.

3.3. Proposed model

Fig. 2 describes the proposed multimodal approach which takes the advantage of 1D convolutional layers as a feature extractor on medical entities. Applying 1D Convolutional Neural Networks (CNN) on text learns the combination of adjacent words and shows successful results for various NLP problems [57]. In the proposed model, K medical entities were extracted from N clinical notes from each patient. These K medical entities are first represented as a sequence of word embeddings with different word representation techniques such as Word2Vec, FastText, and a combination of them. These entities $e_i \in \mathbb{R}^d$ are combined vertically and each patient is represented by a matrix $M \in \mathbb{R}^{k*d}$ where rows are filled with medical entity representations. This patient clinical NER entity matrix (padded where necessary) is represented as:

$$e_{1:k} = e_1 \otimes e_2 \otimes \cdots \otimes e_k \quad (1)$$

where $\otimes$ is the concatenation operator and e refers to the representation of the medical entity and k is the number of the entity. We use a 1D-CNN model similar [58] to extract features from medical entities. We stack three consecutive 1D convolutional layers with filter size 32, 64, and 96. The kernel size is same for three convolutional layers. The output of the last convolutional layer is followed by the max-pooling layer. The final features of the max-pooling layers are concatenated with the features from one layer GRU with 256 hidden units and fed through one fully-connected layer with 512 hidden units.

4. Experimental results

In this section, the results of baseline and multimodal experiments are reported. Moreover, we explain the metrics used for evaluation, and the details about the implementation.

4.1. Setting

Training. For all tasks, we use the patient's first 24 h ICU

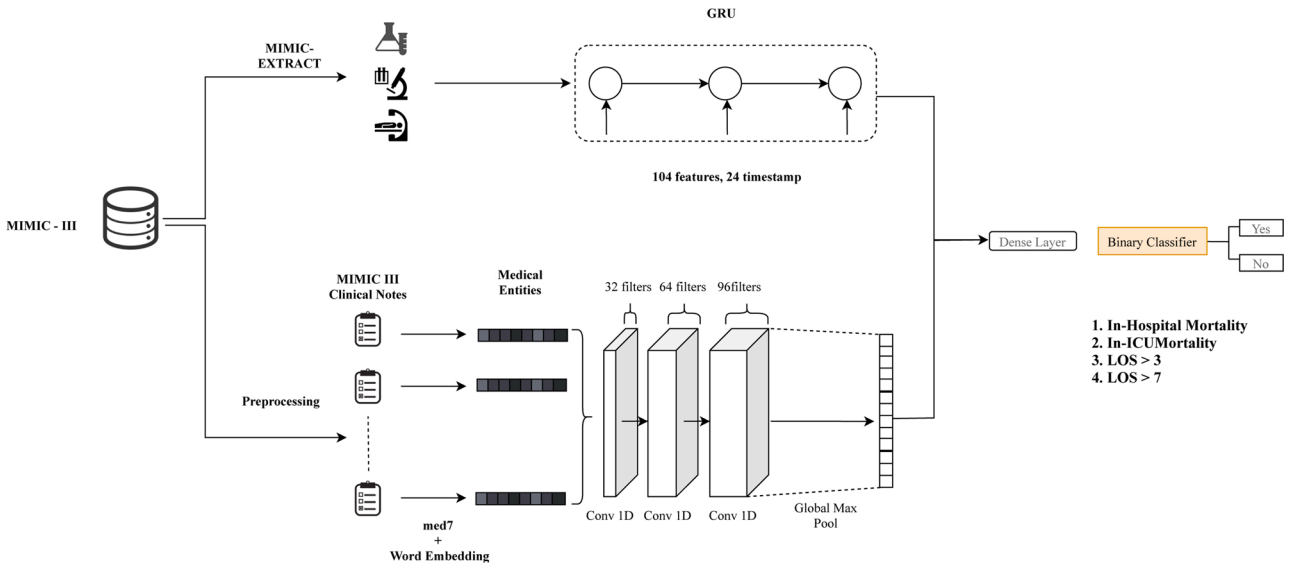

Fig. 2. Overview of Proposed multimodal architecture for predicting the In-Hospital Mortality, In-ICU Mortality, LOS >3, and LOS >7. To extract timeseries features, we use MIMIC-EXTRACT pipeline and fed these features through GRU. We also preprocess the clinical notes and use med7 to extract medical entities. 1D CNN is applied to extract features from medical entity representations. In the final layer, we concatenate features that are extracted from timeseries and medical entities and fed through fully connected layer to predict 4 different binary clinical tasks.

measurements. For multimodal architectures, 0.2 dropout rate is selected at the end of the fully connected layer. A ReLU activation function is used for nonlinearity and L_2 norm for sparsity regularization is selected with the 0.01 scale factor. For the optimization, we use ADAM [59] algorithm with a learning rate of 0.001. All models are trained to minimize the binary cross-entropy loss and we independently tune the hyperparameters – number of hidden layers, hidden units, convolutional filters, filter size, learning rate, dropout rates and regularization parameters on the validation set. Each model is trained for 50 epochs and early stopping is performed on the validation loss. We train each model 10 times with different initialization seeds and report the average performance.

Evaluation metrics. The clinical problems that we work on suffer from class imbalance problem. We use three different metrics which are AUROC, AUPRC and F1. AUROC is a popular robust metric for imbalanced datasets [60]. The second metric AUPRC does not include the true negatives in the calculation and this approach makes it useful for data with many true negative similar to the dataset in this study. F1 is the final metric which calculates the harmonic mean of precision and recall.

Implementation Details. The aforementioned deep learning algorithms are implemented using Keras [61], which runs Tensorflow [62] on its backend. *med7* is used for extracting clinical related entities from clinical notes. All experiments were performed on a computer with NVIDIA Tesla K80 GPU with 24 GB of VRAM, 378 GB of RAM and Intel Xeon E5 2683 processor. The full code of this work is available at https://github.com/tanlab/ConvolutionMedicalNer.

Table 3
Performance comparison of baseline methods. For all four clinical tasks, we report both AUC, AUPRC and F1 scores and the standard deviations. (Bold indicates best results).

Task	Baseline model	Embedding	AUROC	AUPRC	F1
In-hospital mortality	GRU	–	85.04 ± 0.004	52.15 ± 0.009	42.29 ± 0.016
	Doc2Vec multimodal	Doc2Vec	85.96 ± 0.002	54.17 ± 0.004	**46.60** ± 0.016
		Word2Vec	**86.42** ± 0.004	54.22 ± 0.008	45.42 ± 0.013
	Averaged multimodal	FastText	86.09 ± 0.004	**54.47** ± 0.007	45.50 ± 0.010
		Concat	85.98 ± 0.002	54.19 ± 0.008	45.66 ± 0.021
In-ICU mortality	GRU	–	86.32 ± 0.004	46.51 ± 0.011	36.30 ± 0.026
	Doc2Vec multimodal	Doc2Vec	86.80 ± 0.002	48.22 ± 0.006	41.95 ± 0.017
		Word2Vec	**87.17** ± 0.002	**48.47** ± 0.006	42.30 ± 0.021
	Averaged multimodal	FastText	87.14 ± 0.003	48.36 ± 0.006	**42.91** ± 0.014
		Concat	86.90 ± 0.004	48.28 ± 0.007	40.76 ± 0.022
LOS >3 days	GRU	–	67.40 ± 0.003	60.17 ± 0.005	53.36 ± 0.016
	Doc2Vec multimodal	Doc2Vec	**68.90** ± 0.002	**61.88** ± 0.002	54.32 ± 0.008
		Word2Vec	68.63 ± 0.003	61.81 ± 0.003	54.19 ± 0.012
	Averaged multimodal	FastText	68.55 ± 0.003	61.59 ± 0.003	54.46 ± 0.012
		Concat	68.61 ± 0.003	61.69 ± 0.003	**54.70** ± 0.009
LOS >7 days	GRU	–	70.54 ± 0.004	16.35 ± 0.006	**2.33** ± 0.012
	Doc2Vec multimodal	Doc2Vec	**71.63** ± 0.005	17.22 ± 0.004	1.50 ± 0.007
		Word2Vec	71.59 ± 0.005	**17.91** ± 0.006	1.35 ± 0.008
	Averaged multimodal	FastText	71.31 ± 0.008	17.57 ± 0.007	1.02 ± 0.008
		Concat	71.59 ± 0.007	17.67 ± 0.007	1.37 ± 0.013

4.2. Results

4.2.1. Baseline model results

We work on four different clinical tasks with the patient's first 24 h ICU measurements and medical entities. Table 3 summarizes the overall performance of baseline methods. As seen from results, instead of strong results of time-series GRU model, multimodal approaches improve the performance, as expected. For in-hospital mortality prediction, we see an improvement of %1.5 AUROC, %2.5 AUPRC and %4 F1 score compared to the time-series GRU model. For other mortality prediction task, in-ICU mortality, multimodal approach improve the performance around %2 for AUROC and AUPRC and %7 for F1 score. Multimodal approaches also improve the performance of prediction tasks in LOS problem. Both in LOS >3 and LOS >7, all metrics are improved around %1.5. Time-series GRU model only achieves a better F1 score for LOS >7 problem compared to other models.

To demonstrate the effectiveness of medical entities, averaged and Doc2Vec multimodal methods are trained. The results clearly show that using medical entities with time-series based patient features correspondingly improve the mortality and LOS predictions in both types of representations. We observe that averaged and Doc2Vec's performance do not change significantly. Apart from the fact that these two approaches increase the performance of the models, in order to utilize these medical entities in an efficient way, the experiments are carried out labeling the convolutional based deep multimodal model as the most feasible approach.

4.2.2. Proposed model results

In this section, we compare the results of the proposed model against the best scores taken from baseline models and discuss the efficiency and reliability of the proposed model. All results for the proposed model against best baseline scores are provided in Table 4. As shown in Table 3, multimodal approach improves the performance of predictions tasks over the time-series, however we try to use medical entities more efficiently to improve the prediction of multimodal approaches. Except the F1 score of LOS >7 clinical task, the proposed multimodal architecture robustly outperforms all other baseline models for each task.

We associated the efficiency of proposed model with factors mostly related to the usage of medical entities and convolutional based deep multimodal architecture. Due to the advantages of CNNs in capturing local features, it was adapted to various NLP tasks in the literature [63, 64]. Since the high performance of 1D CNN algorithms on text based data is well known [57], we take advantage of convolutional layers in the proposed model. Considering the results obtained from the experiments, using convolution to extract features on medical entities results in consistently better performance. To test the reliability of all the models, the experiments are repeated 10 times with different initialization and the mean performance scores are reported.

5. Discussion

Table 3 shows that the use of medical entity features improve the prediction performance on all clinical tasks. As shown in Table 3, multimodal baseline modals increase all metrics performance which indicates the benefit of using medical entities for predicting mortality and LOS. These experiments also provide an opportunity to compare the medical entity representation methods. Although there is no certain winner for all tasks, in the baseline models, the results show us for mortality prediction tasks, representing the medical entities with averaging method gives better results. For LOS prediction tasks, representing all medical entities together with Doc2Vec is also successful as averaging method. Furthermore, both scores in Tables 3 and 4 gives a chance to compare the word embedding approaches. We do not observe a significant change in performance between word embedding techniques, however pretrained Word2Vec model generally achieves slightly higher scores (around %0.5) than FastText and experimental concatenated embeddings. Apart from these experiments and comparisons, the main motivation is finding an efficient way to combine time-series features with medical entities. Even though both baseline multimodals improve the prediction results compared to timeseries baseline, to make better feature extraction on medical entities, we want to take the advantage of 1D CNN. We stack three 1D convolution operation to extract the features, and then apply 1D max pooling operation over the time-step to obtain a fixed-length vector. By analyzing the results between the proposed and baseline multimodals, we see that 1D CNN based multimodal approach give better results than the averaging and document based embedding methods. In addition to these trials, we also make experiments by using only medical entity features as another baseline. However, only medical entity baseline give poor results (around less than % 10 for all tasks) compared to the timeseries and multimodal, so we do not report these results.

In the literature, there are several studies concerning mortality and LOS prediction. Purushotham et al. [33] work on in-hospital mortality, LOS, and ICD-9 code group predictions. They propose a pipeline to extract 136 raw and 12 clinical aggregate features from MIMIC-III and use more traditional machine learning techniques to make predictions.

Table 4
Proposed model performance comparison with best baseline model. We select the highest score for each metric and each clinical task from baseline methods. (Bold indicates best results).

Task	Model	Embedding	AUROC	AUPRC	F1
In-hospital mortality	Best baseline	–	86.42 ± 0.004	54.47 ± 0.007	46.60 ± 0.016
	Proposed model	Word2Vec	**87.55** ± 0.003	**55.87** ± 0.008	**47.23** ± 0.014
		FastText	87.15 ± 0.002	55.68 ± 0.005	46.87 ± 0.015
		Concat	86.98 ± 0.003	55.35 ± 0.008	46.38 ± 0.027
In-ICU mortality	Best baseline	–	87.17 ± 0.002	48.47 ± 0.006	42.91 ± 0.014
	Proposed model	Word2Vec	**88.35** ± 0.002	**49.23** ± 0.008	43.02 ± 0.029
		FastText	87.85 ± 0.001	48.78 ± 0.009	**43.09** ± 0.026
		Concat	87.66 ± 0.002	48.74 ± 0.009	42.24 ± 0.027
LOS >3 days	Best baseline	–	68.90 ± 0.002	61.88 ± 0.002	54.70 ± 0.009
	Proposed model	Word2Vec	69.54 ± 0.002	62.68 ± 0.003	55.04 ± 0.012
		FastText	69.61 ± 0.003	62.55 ± 0.003	**55.87** ± 0.017
		Concat	**69.93** ± 0.001	**62.77** ± 0.002	55.82 ± 0.008
LOS >7 days	Best baseline	–	71.63 ± 0.005	17.91 ± 0.006	**2.33** ± 0.012
	Proposed model	Word2Vec	**72.55** ± 0.005	**18.78** ± 0.006	1.58 ± 0.001
		FastText	71.81 ± 0.004	18.01 ± 0.004	1.08 ± 0.008
		Concat	71.92 ± 0.007	18.25 ± 0.006	1.38 ± 0.009

In this study, we use MIMIC-Extract to represent the time-varying features and work with a much larger set of 104 clinical aggregate features. Also, all clinical tasks in this study are formulated as a classification problem, while Purushotham et al. formulate the LOS task as a regression problem and use the mean squared error to evaluate their model performance. Nallabasannagari et al. [65] propose a model that combines multiple data sources in MIMIC-III to predict in-hospital mortality and LOS >7. To extract features from raw data, they use the same strategy for each data source. Features containing free-text data are split on whitespace to create tokens, and other features are combined with them. This tokenization process creates a two-dimensional array which consists of hospital admissions and a list of tokens for each hospital admission. The architecture of the model is simply constructed with embedding layers, an averaging layer, and multiple dense layers. With this proposed method, they avoid discarding too much patient data. However, the complex and multi-source raw patient data may need more detailed preprocessing steps. In addition, in our study, instead of using only dense layers, to benefit from temporal information in the data, we apply time-series based algorithms to the patient's time-varying features and use 1D CNN for text based medical entity data with different word embedding techniques. Another work, Jin. et al. [50] propose a multimodal neural network that uses time series features with unstructured clinical notes and tries to predict in-hospital mortality risk for ICU patients. To represent clinical notes, only Document Vector through Corruption (Doc2VecC) is used. In this study, we work with three more additional outcomes which are in-ICU mortality, LOS >3, and LOS >7 rather than just in-hospital mortality. We also compare different types of methods to represent medical entities like averaging, Doc2Vec, and convolutional based. Therefore, to discuss the effect of different word embedding methods, all experiments are carried out with Word2Vec, FastText, and concatenation of them.

6. Conclusion and future work

Over the past decade, there has been increased attention to improve mortality and LOS prediction performance. Predicting any complications and saving patient's life is an important task for healthcare system which motivates us to work on mortality prediction. LOS is another important clinical problem to improve hospital performance and better healthcare resource utilisation. In this work, we present 1D-CNN based multimodal deep learning architecture that use time-series features and medical entities together and this model outperforms several baselines. The proposed model performance gain over multimodal baselines is around %1–%1.5 AUPRC, and the improvement over time-series baseline is around %2.5–%3 AUPRC. We also make experiments to investigate the effect of different word embedding algorithms to solve clinical problems and report the results.

Despite these contributions, the proposed model also has some limitations that can be addressed in a future work. First, we only use context-independent word embedding techniques such as Word2Vec, FastText to represent medical entities. Second, the pre-trained models that are used in this study such as word embeddings and clinical NER models are trained on MIMIC-III dataset. Moreover, the time-series feature extraction pipeline is also specific to MIMIC-III dataset. Therefore, using the proposed methodology on an EHR dataset different from MIMIC-III may lead to some problems. Third, although the proposed model is shown to be successful on clinical tasks, the level of explainability of the predictions are low.

This work can be extended in multiple directions. First, recent context-dependent embeddings like BERT to represent medical entities can be utilized. Second, clinical word embeddings and clinical NER model can be trained on a more general clinical based corpus rather than MIMIC-III. Third, we can add more features associated with patients such as prescription data and diagnosis codes to improve the prediction performance. Another thing we may consider in the future is to use more advanced deep learning architectures with attention mechanism to improve explainability and the accuracy of predictions.

Funding

This study has been partially funded by The Scientific and Technological Research Council of Turkey (TUBITAK), Grant Number:120E173.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. Mimic-iii, a freely accessible critical care database. Sci Data 2016;3:160035.
[2] Ghassemi M, Wu M, Hughes MC, Szolovits P, Doshi-Velez F. Predicting intervention onset in the icu with switching state space models. AMIA Summits Trans Sci Proc 2017;2017:82.
[3] McDermott MB, Yan T, Naumann T, Hunt N, Suresh H, Szolovits P, et al. Semi-supervised biomedical translation with cycle wasserstein regression gans. Thirty-second AAAI conference on artificial intelligence 2018.
[4] Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 2019;109:79–84.
[5] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. Machine learning for healthcare conference 2016:301–18.
[6] Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. Advances in neural information processing systems. 2016. p. 3504–12.
[7] Caballero Barajas KL, Akella R. Dynamically modeling patient's health state from electronic medical records: a time series approach. Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining 2015:69–78.
[8] Song H, Rajan D, Thiagarajan JJ, Spanias A. Attend and diagnose: clinical time series analysis using attention models. Thirty-second AAAI conference on artificial intelligence 2018.
[9] Suresh H, Gong JJ, Guttag JV. Learning tasks for multitask learning: heterogenous patient populations in the ICU. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining 2018:802–10.
[10] Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. 2018 (arXiv preprint), arXiv:1802.05695.
[11] Boag W, Doss D, Naumann T, Szolovits P. What's in a note? Unpacking predictive value in clinical note representations. AMIA Summits Trans Sci Proc 2018;2018:26.
[12] Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The stanford corenlp natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations 2014:55–60.
[13] Honnibal M, Johnson M. An improved non-monotonic transition system for dependency parsing. Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics; 2015. p. 1373–8. https://aclweb.org/anthology/D/D15/D15-1162.
[14] Gasmi H, Bouras A, Laval J. LSTM recurrent neural networks for cybersecurity named entity recognition. ICSEA 2018;11:2018.
[15] Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: a transferable clinical natural language processing model for electronic health records. 2020 (arXiv preprint), arXiv:2003.01271.
[16] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013 (arXiv preprint), arXiv:1301.3781.
[17] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. 2016 (arXiv preprint), arXiv:1607.01759.
[18] Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. 2015 (arXiv preprint), arXiv:1511.03677.
[19] Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017;24(2):361–70.
[20] Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232–43.
[21] Dybowski R, Gant V, Weller P, Chang R. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 1996;347(9009):1146–50.
[22] Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: customized mortality prediction. J Pers Med 2012;2(4):138–48.
[23] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. Apache-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981;9(8):591–7.
[24] Le Gall J-R, Lemeshow S, Saulnier F. A new simplified acute physiology score (saps ii) based on a European/North American multicenter study. JAMA 1993;270(24):2957–63.
[25] Vincent J-L, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996;22:707–10.

[26] Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform 2017;108:185–95.
[27] Sadeghi R, Banerjee T, Romine W. Early hospital mortality prediction using vital signals. Smart Health 2018;9:265–74.
[28] Darabi HR, Tsinis D, Zecchini K, Whitcomb WF, Liss A. Forecasting mortality risk for patients admitted to intensive care units using machine learning. Proc Comput Sci 2018;140:306–13.
[29] Yakovlev A, Metsker O, Kovalchuk S, Bologova E. Prediction of in-hospital mortality and length of stay in acute coronary syndrome patients using machine-learning methods. J Am Coll Cardiol 2018;71(11 Supplement):A242.
[30] Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: a survey. Health Serv Manag Res 2017;30(2):105–20.
[31] Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019;6(1):1–18.
[32] Wang S, McDermott MB, Chauhan G, Ghassemi M, Hughes MC, Naumann T. Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. Proceedings of the ACM conference on health, inference, and learning 2020:222–35.
[33] Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018;83:112–34.
[34] Si Y, Roberts K. Deep patient representation of clinical notes via multi-task learning for mortality prediction. AMIA Summits Transl Sci Proc 2019;2019:779.
[35] Liu J, Zhang Z, Razavian N. Deep ehr: chronic disease prediction using medical notes. 2018 (arXiv preprint), arXiv:1808.04928.
[36] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018 (arXiv preprint), arXiv: 1810.04805.
[37] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019. p. 5754–64.
[38] Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. 2019 (arXiv preprint), arXiv:1904.05342.
[39] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical bert embeddings. 2019 (arXiv preprint), arXiv:1904.03323.
[40] Zhu H, Paschalidis IC, Tahmasebi A. Clinical concept extraction with contextual word embedding. 2018 (arXiv preprint), arXiv:1810.10566.
[41] Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend medical: a named entity recognition and relationship extraction web service. 2019 (arXiv preprint), arXiv:1910.07419.
[42] Fraser KC, Nejadgholi I, De Bruijn B, Li M, LaPlante A, Abidine KZE. Extracting umls concepts from medical text using general and domain-specific deep learning models. 2019 (arXiv preprint), arXiv:1910.01274.
[43] Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020;27(3): 457–70.
[44] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. ICML 2011.
[45] Karpathy A, Joulin A, Fei-Fei LF. Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems. 2014. p. 1889–97.
[46] Ilievski I, Feng J. Multimodal learning and reasoning for visual question answering. Advances in neural information processing systems. 2017. p. 551–62.
[47] Mroueh Y, Marcheret E, Goel V. Deep multimodal learning for audio-visual speech recognition. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2015. p. 2130–4.
[48] Khadanga S, Aggarwal K, Joty S, Srivastava J. Using clinical notes with time series data for ICU management. 2019 (arXiv preprint), arXiv:1909.09702.
[49] Shukla SN, Marlin BM. Integrating physiological time series and clinical notes with deep learning for improved ICU mortality prediction. 2020 (arXiv preprint), arXiv: 2003.11059.
[50] Jin M, Bahadori MT, Colak A, Bhatia P, Celikkaya B, Bhakta R, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. 2018 (arXiv preprint), arXiv:1811.12276.
[51] Chen M. Efficient vector representation for documents through corruption. 2017 (arXiv preprint), arXiv:1707.02377.
[52] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8): 1735–80.
[53] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014 (arXiv preprint), arXiv:1412.3555.
[54] Neumann M, King D, Beltagy I, Ammar W. Scispacy: fast and robust models for biomedical natural language processing. 2019 (arXiv preprint), arXiv:1902.07669.
[55] Mulyar A, Mahendran D, Maffey L, Olex A, Matteo G, Dill N, et al. Tac srie 2018: extracting systematic review information with medacy. Strain 2020;372:338.
[56] Le Q, Mikolov T. Distributed representations of sentences and documents. International conference on machine learning 2014:1188–96.
[57] Kim Y. Convolutional neural networks for sentence classification. 2014 (arXiv preprint), arXiv:1408.5882.
[58] Öztürk H, Özgür A, Ozkirimli E. Deepdta: deep drug-target binding affinity prediction. Bioinformatics 2018;34(17):i821–9.
[59] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014 (arXiv preprint), arXiv:1412.6980.
[60] Davis J, Goadrich M. The relationship between precision-recall and roc curves. Proceedings of the 23rd international conference on machine learning 2006: 233–40.
[61] Chollet F. Keras. 2015. https://github.com/fchollet/keras.
[62] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous systems, software available from tensorflow.org. 2015. http://tensorflow.org/.
[63] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. 2014 (arXiv preprint), arXiv:1404.2188.
[64] Ma M, Huang L, Xiang B, Zhou B. Dependency-based convolutional neural networks for sentence embedding. 2015 (arXiv preprint), arXiv:1507.01839.
[65] Reddy Nallabasannagari A, Reddiboina M, Seltzer R, Zeffiro T, Sharma A, Bhandari M. All data inclusive, deep learning models to predict critical events in the medical information mart for intensive care iii database (mimic iii). 2009. arXiv-2009 (arXiv e-prints).