2019 IEEE 35th International Conference on Data Engineering (ICDE)

Knowledge-Aware Deep Dual Networks for Text-Based Mortality Prediction

Ning Liu[†], Pan Lu[†], Wei Zhang[§]*, and Jianyong Wang[†‡]
[†]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[‡]Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University, Xuzhou, China
Email: victorliucs@gmail.com, lupantech@gmail.com, jianyong@tsinghua.edu.cn
[§]School of Computer Science and Software Engineering, East China Normal University, Shanghai, China
Email: zhangwei.thu2011@gmail.com

***Abstract*—Mortality prediction is one of the essential tasks in medical data mining and is significant for inferring clinical outcomes. With a large number of medical notes collected from hospitals, there is an urgent need for developing effective models for predicting mortality based on them. In contrast to structured electronic health records, medical notes are unstructured texts written by experienced caregivers and contain more complicated information about patients, posing more challenges for modeling. Most previous studies rely on tedious hand-crafted features or generating indirect features based on some statistical models such as topic modeling, which might incur information loss for later model training. Recently, some deep models have been proposed to unify the stages of feature construction and model training. However, domain concept knowledge has been neglected, which is important to gain a better understanding of medical notes. To address the above issues, we propose novel Knowledge-aware Deep Dual Networks (K-DDN) for the text-based mortality prediction task. Specifically, a simple deep dual network is first proposed to fuse the representations of medical knowledge and raw text for prediction. Afterward, we incorporate a co-attention mechanism into the basic model, guiding the knowledge and text representation learning with the help of each other. Experimental results on two publicly real-world datasets show the proposed deep dual networks outperform state-of-the-art methods and the co-attention mechanism can further improve the performance.**

***Keywords*-mortality prediction, deep learning, medical concept**

I. Introduction

Mortality prediction, aiming at providing an accurate assessment of the risk of death, is one of the essential tasks in the field of medical data mining. For example, clinicians in intensive care units (ICU) can flexibly take appropriate decisions based on the estimated mortalities, which is promising to improve the effective health management and decrease the number of patients who will suffer from prevented deaths [1]. In addition, it is also important for the hospitals with quite limited resources, such as the lack of clinicians. Due to the great value of mortality prediction, a growing body of studies [2] have investigated this task from various aspects, such as how to model different measurements [3], [4] and how to capture the dependency of related time series data [5], [6].

With the development of electronic health records, medical notes have been largely collected and some of them are freely available for researchers, initiating a new research direction for applying text mining techniques to medical data mining. In contrast to standard medical examinations which usually are structured data, medical notes are unstructured data written by experienced caregivers, contain more detailed information about patients, and can support further treatment decisions [7]. Therefore, modeling medical notes becomes a potential solution to accurately reveal the patient's state for modeling mortality prediction and some pioneering studies [7]–[11] have made attempts in this regard.

A commonly adopted paradigm in majority of these studies [7]–[10] is to first construct hand-crafted features or generated features based on some statistical models. For example, latent Dirichlet allocation (LDA) [12] is leveraged to generate topic distributions for different notes, which could be regarded as one type of features for mortality prediction. After extracting these features, classical classification models (e.g., support vector machine [13]) are then trained to generate predictions. However, we argue that the above paradigm is sub-optimal since the feature construction and the model training are separated into two independent stages. Since the feature construction stage is not directly guided by the model optimization, it might incur information loss for later prediction.

To bridge this gap, deep learning models are exploited for the task to unify the two stages [11]. Nevertheless, medical concepts have been neglected in these models, which may degrade the prediction performance. It is because that medical concepts contain knowledge from professional domain-specific knowledge bases and summarize corresponding notes from a high level. In domain-specific text classification tasks, like mortality prediction based on medical notes, it is important to consider knowledge information brought from domain concepts [14]. For example, the medical notes we used in the experiments have a sentence "there is no mediastinal vascular engorgement to suggest cardiac tamponade". If only word-level information in the original text is used, then concept-level information may not be captured since the holistic concept "cardiac tamponade" could be represented by "cardiac" and "tamponade" separately. In summary, how to effectively

*Corresponding author.

2375-026X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDE.2019.00127

model medical notes for mortality prediction still remains an unresolved challenge.

To tackle the above challenge, we propose two novel Knowledge-aware Deep Dual Networks (K-DDN), which not only unify the stages of feature construction and model training, but also take medical concepts into modeling. Specifically, we first introduce a Basic Knowledge-aware Deep Dual Network (BK-DDN) which consists of two neural network branches. One of the branches takes the original raw text as input and obtains its representation through convolutional neural networks (CNN), inspired by [15]. Simultaneously, the other one regards medical concepts as input and also learns the corresponding representation by CNN. BK-DDN further fuses the two representations into an integrated representation which is later fed into higher layers for classification. To further improve the performance, we propose a co-attention mechanism and incorporate it into the BK-DDN model, named as Advanced Knowledge-aware Deep Dual Network (AK-DDN). The basic motivation of the mechanism is to interact with the representation learning of concept with that of raw text, hoping to guide the learning with the help of each other.

The main contributions of this paper are summarized as follows:

- We extract the Unified Medical Language System (UMLS) concepts from the medical notes and address to learn both the word-level and concept-level representations of the medical notes for the mortality prediction task.
- We first propose a Basic Knowledge-aware Deep Dual Network (BK-DDN) model to unify the representation learning and model training stages. To our knowledge, this is the first study to investigate the power of combining the domain-related knowledge with raw text in an end-to-end deep learning framework for medical note classification.
- We further incorporate the novel co-attention mechanism into the basic model, achieving the mutual learning of word-level representation and concept-level representation.
- We conduct comprehensive experiments on two real-world medical datasets with reasonable measurements and the results show our model outperforms the state-of-art methods.

II. Related Work

In this section, we briefly summarize some related studies from the following aspects: text classification, mortality prediction, knowledge-aware medical data mining, and attention mechanism.

A. Text Classification

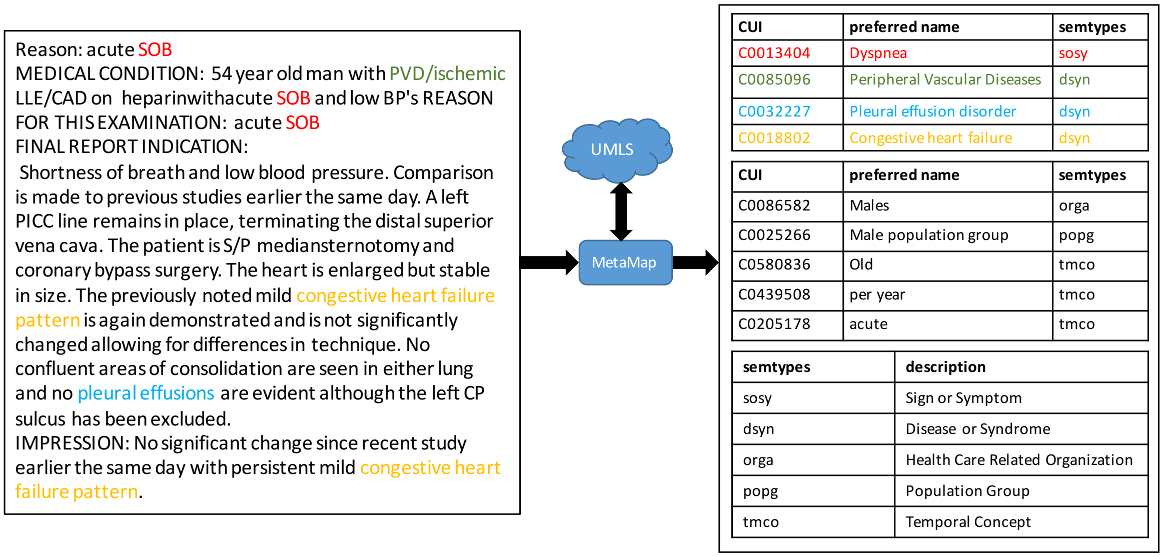
Document classification has been extensively studied due to its wide applications, such as sentiment classification, medical diagnosis, spam filtering, information retrieval, etc. Many standard classification models can be adapted to this task. Specifically, they regard words as basic feature units and rely on human designed features. Manevitz et al. proposed a variant of support vector machine for one-class classification by using different features, including term frequency representation and term frequency-inverse document frequency (TF-IDF) [16]. In [17], support vector machines and navies Bayes are used with word-level features as well. Forman [18] investigated feature selection methods to determine whether each word contributes to the classification performance. In addition to the word features, some other studies adopt POS tagging or even more complex tree kernel based features [19].

In the past few years, deep learning has been widely used in document classification and achieved remarkable success. In [15], convolutional neural networks are used for sentence classification. Wang et al. [14] used convolutional neural networks which can combine knowledge and character level features for classifying short text. Despite convolutional networks, recurrent neural networks [20] and recursive neural networks [21] also find their applications in text classification. In this paper, we leverage convolutional neural networks as cornerstones to construct our models, for their simplicity and efficiency.

B. Mortality Prediction

Mortality prediction is significant for inferring clinical outcomes. Some studies have made use of structural features in medical electronic records and generated scores of illness severity. These features involve demographics, medications, and laboratory tests [1]. Early approaches are mainly rule based methods, such as APACHE [22], SAPS-II [23], and SOFA [24]. Recent models utilize the structured features from the perspective of multi-task learning [1], imbalance learning [25], and time-series learning [26]. Since the focus in this paper is to effectively model medical notes, we do not test these models in the experiments. However, these studies are complementary to our study.

For the studies of using medical notes for mortality prediction, both Saria et al. [27] and Lehman et al. [9] considered to utilize the concepts extracted from the notes and modeled them by logistic regression and LDA, respectively. Ghassemi et al. [7] adopted LDA to model the raw text of medical notes and train the SVM model for prediction. Jo et al. [28] further modeled the temporal dynamics between the nursing notes by the proposed state transition topic model and SVM was adopted as well. However, as we mentioned in Section I, the above paradigm is suboptimal since the feature construction and model training are separated into two independent stages. To address this issue, the deep learning model based on convolutional neural network was proposed to unify the two stages [11]. Unfortunately, medical concepts have been overlooked in this paper, which might influence the prediction performance. In contrast to existing relevant models, we develop knowledge-aware neural dual networks, hoping to not only fuse the feature learning and model training but also incorporate medical concepts and capture their knowledge through deep learning methodologies.



Reason: acute SOB
MEDICAL CONDITION: 54 year old man with PVD/ischemic
LLE/CAD on heparinwithacute SOB and low BP's REASON
FOR THIS EXAMINATION: acute SOB
FINAL REPORT INDICATION:
Shortness of breath and low blood pressure. Comparison
is made to previous studies earlier the same day. A left
PICC line remains in place, terminating the distal superior
vena cava. The patient is S/P mediansternotomy and
coronary bypass surgery. The heart is enlarged but stable
in size. The previously noted mild congestive heart failure
pattern is again demonstrated and is not significantly
changed allowing for differences in technique. No
confluent areas of consolidation are seen in either lung
and no pleural effusions are evident although the left CP
sulcus has been excluded.
IMPRESSION: No significant change since recent study
earlier the same day with persistent mild congestive heart
failure pattern.

CUI	preferred name	semtypes
C0013404	Dyspnea	sosy
C0085096	Peripheral Vascular Diseases	dsyn
C0032227	Pleural effusion disorder	dsyn
C0018802	Congestive heart failure	dsyn

CUI	preferred name	semtypes
C0086582	Males	orga
C0025266	Male population group	popg
C0580836	Old	tmco
C0439508	per year	tmco
C0205178	acute	tmco

semtypes	description
sosy	Sign or Symptom
dsyn	Disease or Syndrome
orga	Health Care Related Organization
popg	Population Group
tmco	Temporal Concept

Fig. 1: example of the medical notes of a patient: the left part is a demo of a note from a patient and the right part is the umls concepts extracted from the note with detailed information and the upper and middle tables depict the umls concepts of the note and the lower table depicts the descriptions of semantic meanings of the concepts. Concepts in the upper table marked with colors is the concepts related to diseases and those in the middle table are concepts of general meanings which can be filtered out by their semantic meanings. Both of the note and the concepts are truncated as the space is limited.

C. Knowledge-aware Medical Data Mining

Some studies have considered the medical knowledge in the domain of medical data mining. Wang et al. [29] utilized the knowledge of phenotypes to regularize their representation learning in matrix factorization and their experiments show it can improve the capability of discovering new phenotypes. Zhang et al. [30] leveraged the prior knowledge of adverse drug interaction in the reinforcement learning framework to control the recommendation of medication lists. Cao et al. [31] proposed to guide the neural network learning with the help of the knowledge from domain concept dictionaries. However, there are major differences between ours and Cao et al. [31]. Firstly, Cao et al's work is based on short text classification which is not our paper's settings. Secondly, they explored a word-replace strategy to generate text corpus which may lose information while our's work models both words and concepts at the same time. More recently, Wang et al. [32] combined the knowledge between diseases with other patient body measurements for medication recommendation. However, most of the existing studies do not consider to integrate neural network learning with medical knowledge for the mortality prediction task, which is the focus of this paper.

D. Attention Mechanism

The attention mechanism aims to learn different importance proportions for each term to be considered. Since being put forward by for machine translation [33], it has gained enormous attention in diverse applications such as image captioning [34], question answering [35], popularity prediction [36], etc. With regard to the field of medical data mining, the aforementioned studies [30]–[32] all exploited the advantages of the attention mechanism. Inspired by these studies, we incorporate the attention mechanism into our deep dual networks to benefit the mutual representation learning of medical concepts and raw text.

III. Preliminary

A. Problem Definition

In general, document classification is the task of assigning a category label $y_i (y_i \in Y)$ to a given document $d_i (d_i \in D)$, where D is a document set and Y is a category set. Formally, the solutions to this task are to learn a mapping function ϕ which is defined as follows:

$$\phi : d_i \longrightarrow y_i$$

In contrast to the above general document classification task, domain specific document classification task, such as the medical text based mortality prediction problem focused on in this paper, needs to additionally consider domain knowledge to gain satisfied prediction performance. To this end, we extend the problem setting of general document classification to the medical text based mortality prediction problem as the following.

Assume C is a semantic concept set, D represents a medical document set, and each patient is uniquely associated with a medical document. For a given document $d_i \in D$, we denote its corresponding concepts as $c_i \subset C$ the major aim of the *problem is to learn a following mapping function:*

$$\phi :< d_i, c_i > \longrightarrow y_i$$

where $y_i \in Y = \{alive, dead\}$.

Formally, we denote $c_i = \{c_{i,1}, c_{i,2}, ..., c_{i,m_i^c}\}$ as the concepts extracted from the medical document d_i, and $d_i = \{w_{i,1}, w_{i,2}, ..., w_{i,m_i^w}\}$ where m_i^c is the number of concepts in the document and m_i^w is the number of words. Following [7], we address the above problem in this paper by predicting whether a patient will die in hospital, within 30 days or one year after discharge from the hospital.

B. Unified Medical Language System

Unified Medical Language System (UMLS) [37], composed of a set of standard medical documents and softwares, is developed by National Library of Medicine (NLH)[1] to standardize and improve the medicine related research.

The UMLS concepts in the medical notes can be extracted using Metamap [38], an efficient tool developed by NLH to map biomedical texts to the UMLS Metathesaurus. By using natural language processing and computational technologies, Metamap is used as a powerful tool in medical text mining [39], [40], automatic disease prediction systems [41], clinical trial analysis [42], etc. The use of Metamap on raw medical texts reveals domain-specific knowledge and gives a higher abstraction of medical texts than raw texts. Figure 1 illustrates a simple example of medical texts and UMLS Metathesaurus extracted by Metamap.

C. Query Based Attention

Attention mechanism is widely used in various neural networks and has been successfully applied into many areas such as machine translation, speech recognition and sentiment analysis. Given a query vector q, key-value pairs $(k_j, v_j)_{j=1}^{j=N}$, the function of attention is to map a tuple $(q, (k_j, v_j))$ to an output. The output of attention is computed by a weighted-sum operation, which is composed of the query q and the key k_j. Therefore, an attention function is defined as:

$$ATTEN(q, k, v) = softmax(q \cdot k^T)v$$

where the softmax function is defined as:

$$softmax(q \cdot k^T)_j = \frac{e^{q \cdot k_j^T}}{\sum_{j'} e^{q \cdot k_{j'}^T}}, j = 1, 2, \cdots, N$$

Attention mechanism can be considered as a weighted sum of values given a query. The weights generated can be visualized to show the importance of each value.

IV. Basic Knowledge-aware Deep Dual Networks

The Basic Knowledge-aware Deep Dual Networks (BK-DDN) first proposed has two main components. One is a neural network that transforms a text into a text representation vector and the other involves concept modeling to transform concepts of the text into a vector representation.

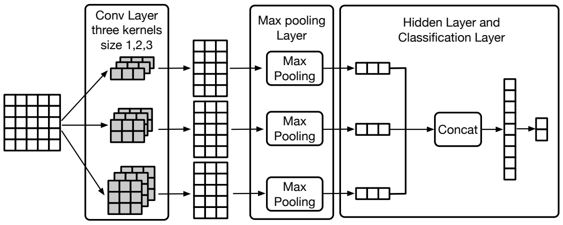


Fig. 2: description of word-level note representation(Text CNN | Concept CNN) using three convolutional layers, three max pooling layers and a hidden layer.

A. Words and Concepts embedding

Different from the previous work using static pre-trained word embedding matrix, we train the concept and word embeddings while making predictions. Firstly, the preferred name of concepts may be composed of more than one word and each concept may have more than one alias, which make it difficult for representing concepts via pre-trained word embeddings. Secondly, the static pre-trained word embeddings are usually based on general corpus such as Twitter, Wikipedia, etc. Using such static pre-trained word embeddings may introduce more noises as the probabilities of domain-related words in general corpus are unlike those in medical documents. And in the general corpus, some of the medical terms are not common words which may not exist or occur with low frequency, thus these terms are likely to be filtered out before word embedding training. Finally, the general meaning of words may be different from the meaning of medical texts. Therefore, in order to overcome those drawbacks of using static word embeddings, we learn the embedding matrix during model training.

Formally, we denote the word embedding matrix as $E^w \in \mathbb{R}^{n^w \times l^w}$, where n^w is the number of unique words in the medical notes and l^w is the dimension of word embedding vector. Similarly, we denote the concept embedding matrix $E^c \in \mathbb{R}^{n^c \times l^c}$ where n^c is size of vocabulary extracted from the medical notes and l^c is the dimension of the embedding vector of a concept.

B. Word-level Text Representation

Previous work [15] shows that convolutional neural networks are powerful in text classification. Inspired by this, the component has eight layers: one input layer, three convolutional layers, three pooling layers, and one hidden layer.

1) Input Layer: Similar to traditional embedding process, the input layer is designed to transform the word vectors of the document d_i into its embedding matrix $W_i \in \mathbb{R}^{m_i^w \times l^w}$. For all words in the document, we first obtain their one-hot representation $O_i^w \in \{0, 1\}^{m_i^w \times n^w}$ with only one dimension having a value 1 in each row, indicating the corresponding word occurring in the current position. Then the word embedding matrix W_i can be generated by:

$$W_i = O_i^w \cdot E^w$$

[1] https://www.nlm.nih.gov/

2) Convolutional Layer: The function of the convolutional layers is to extract higher level word features from the embedding matrix of a document. Similar to previous work, we vary the size of filters to achieve modeling different granularities. Specifically, we set three filters $f^h \in R^{h \times l^w}$ with $h = 1, 2, 3$. Thus we can acquire unigram, bigram, and trigram information of the word embedding matrix W_i. After convolution operation, three feature maps denoted as w_i^1, w_i^2, w_i^3 are generated by:

$$w_{ij}^k = g(f^k \cdot [W_{i,j:j+k,:}^w] + b), k = 1, 2, 3$$

where g is a activation function, and b is a bias vector. Here we use *RELU* as the activation function, which is defined by:

$$RELU(x) = max(0, x)$$

3) Max Pooling Layer: The pooling layer is used to select the significant hidden features that are obtained from the convolution layer. In our network, we apply max pooling operation on each feature map generated from convolution layer to choose the highest score in each column and each feature map generates a fixed length vector. Therefore, these vectors generated by max pooling operation contain the most important word level information of the text. In details, each convolutional unit needs a max pooling operation. Thus, three max pooling layers are used to extract important unigram, bigram, and trigram feature maps. In formal, the outputs denoted as p_i^1, p_i^2, p_i^3 of Max Pooling layer can be computed by:

$$p_{ij}^k = max(w_{i,:,j}^k), k = 1, 2, 3$$

where $w_{i,:,j}^k$ denotes the j^{th} column of feature map f^k.

4) Hidden Layer: The function of the hidden layer is to get a word-level representation of the text by combining the unigram, bigram and trigram vectors denoted as p_i^1, p_i^2, p_i^3 from the max pooling layer. In our model, we concatenate($\oplus$) the three vectors from the previous layers and get a final representation. And the word-level representation of the medical note can be computed by:

$$v_i^w = p_i^1 \oplus p_i^2 \oplus p_i^3$$

C. Concept-level Text Representation

Similar to text modeling in Section IV-B, this component takes the concepts of a text as input and outputs a fixed size of vector which contains information on the concept-level information of a text. This sub component is composed of eight layers: one input layer, three convolutional layers, three max pooling layers and a hidden layer. The input layer takes the concept one-hot vectors $O_i^c \in \mathbb{R}^{m_i^c \times n^c}$ of a text as the input and outputs a concept embedding matrix $C_i \in \gtrdot^{\mathbb{C}}_{\beth} \times \lessdot$ of the medical note. And the convolutional layers with filters $f^c \in R^{k^c \times l^c}$ of size 1, 2, 3 take the concept embedding matrix as the input and generates three concept-level feature maps $\tilde{w}_i^1$, $\tilde{w}_i^2$, $\tilde{w}_i^3$. After that, a max pooling operation over columns of the feature map is used to generate three vectors which contain concept-level important information. Then a hidden layer is used to concatenate the concept-level vectors and generates the final concept-level representation of the concepts denoted as v_i^c. Therefore, v_i^c can be computed by:

$$C_i = O_i^c \cdot E^c$$

$$\tilde{w}_i^k = RELU(f_i^c \cdot [C_{i,j:j+k,:}] + b), k = 1, 2, 3$$

$$\tilde{p}_{ij}^k = max(\tilde{w}_{i,:,j}^k), k = 1, 2, 3$$

$$v_i^c = \tilde{p}^1 \oplus \tilde{p}^2 \oplus \tilde{p}^3$$

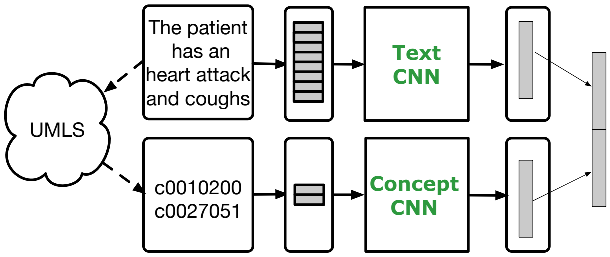


Fig. 3: description of BK-DDN: this figure gives a overview of architecture of BK-DDN, which is composed of two main components: the upper is a Text CNN for word-level representation generation and the lower is a Concept CNN for concept-level representation generation.

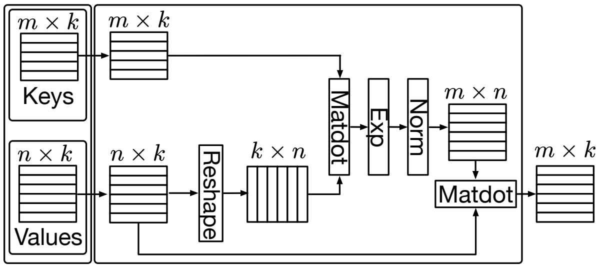


Fig. 4: description of attention based interaction (ATTI): The component takes two matrix (keys and values) as its input and generates a weighted sum of values corresponding to the keys.

With the representation of the medical note at both word-level and concept-level aspects denoted as v_i^w and v_i^c, we generate the final representation of the medical note of a patient by concatenating the two representations. After that, we use a dense layer to shrink the dimension of the final output of the model via a softmax function:

$$output_i = softmax(\Theta \cdot (v_i^w \oplus v_i^c) + b)$$

where $\Theta \in R^{2 \times |v_i^w \oplus v_i^c|}$ and b is the bias vector of the dense layer. The basic Text CNN and Concept CNN are illustrated in Figure 2. The overall structure of our BK-DDN is described in Figure 3.

V. Advanced Knowledge-Aware Deep Dual Networks

In our proposed BK-DDN in Section IV, the word-level and concept-level representations of a document are trained in parallel. Therefore, the upper component and the lower component have no interaction with each other. Note that the concepts extracted from the text are triggered by some words in the notes. The structure of BK-DDN is lack of information because it only extracts concept-level information without the consideration of the interactions. Indeed, concepts and words are highly related as the process of concept extraction is based on the words in the document so that the interactions between words and concepts contain more information than modeling concepts and words independently. Therefore, adding word information in concept-level representation and adding concept information in word-level representation can learn better concept-level and word-level information of a document.

In order to make full use of words and concepts in a medical note, we propose a novel model to consider the interaction between words and concepts. We propose Advanced Knowledge-aware Deep Dual Networks (AK-DDN). It utilizes attention mechanism to generate word-based interaction with concepts and concept-based interaction with words. We leverage the information stored in the word and concept embeddings to model interactions between concepts and words. The input of the model is the same as the BK-DDN and we append two main components: one is called Word-based Interaction with Concepts and the other is called Concept-based Interaction with Words.

1) Concepts-based interaction with Words: Recall that in the BK-DDN, we get a concept-level representation via the convolutional layer, the max pooling layer, and the hidden layer. In this section, we further use the word information to guide the learning process of concept-level representation, we use an attention based interaction method (ATTI), which is illustrated in Figure 4.

Given a word embedding matrix $W_i \in \mathbb{R}^{m_i^w \times l^w}$ and a concept embedding matrix $C_i \in \mathbb{R}^{m_i^c \times l^c}$, we generate a concepts-based interaction using each word embedding W_{ij} as query and the concept embedding matrix C_i as the values. Then the word-based interaction with concepts matrix I_i^c can be generated by:

$$\omega_{ij}^w = softmax(W_{ij} \cdot (C_i)^T)$$

$$I_{ij}^c = \sum_k^{m_i^c} \omega_{ijk}^w \times C_{ik}$$

Using each word as a query to generate concepts-based interaction $I_i^c \in \mathbb{R}^{m_i^w \times l^c}$, the interaction matrix can contain words-related information, and each row of the matrix can be considered as a weighted sum of concept embedding related to the word. Therefore, with the information from words, each row of the final interaction matrix represents the combination information of related concepts towards the word.

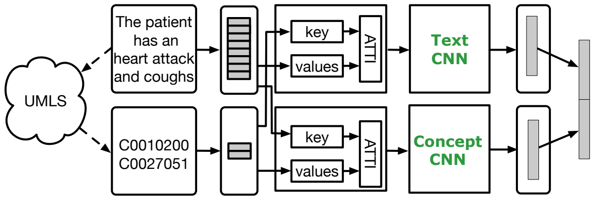


Fig. 5: description of AK-DDN: this figure gives a overview of the architecture of AK-DDN. Different from the previous knowledge aware model, we use both word based interaction and concept based interaction before modeling word-level representation and concept-level representation using attention based interaction.

2) Words-based interaction with Concepts: Similar to the process of generating concepts-based interaction, we generate a word-based interaction I_i^w by taking each concept embedding $C_{ij} \in \mathbb{R}^{l^c}$ as the query and the words embedding matrix $W_i^w \in \mathbb{R}^{m_i^w \times l^w}$ as the values. Therefore, the interaction matrix I_i^w can be generated via the following:

$$\omega_{ij}^c = softmax(C_{ij} \cdot {W_i}^T)$$

$$I_{ij}^w = \sum_k^{m_i^w} \omega_{ijk}^c \times W_{ik}$$

Similar to concept-based interaction in Section V-1, the concept information are combined into the generated interaction matrix $I_i^w \in \mathbb{R}^{m_i^c \times l^w}$, where each row of the matrix can be considered as the weighted sum of the word embeddings with respect to the concept.

After generation of the concept-based interaction I_i^c and the word-based interaction I_i^w, we use two separate convolutional neural networks for modeling both of the concept-based and the word-based interactions. The overall structure of our AK-DDN model is illustrated in Figure 5.

VI. Model Optimization

We denote all parameters of our network as θ, and concepts extracted from the medical notes as C and their corresponding context words as W. Here we denote the training data as $X = \{C, W\}$, and the set of class labels as Y. For a given input $x \in X$, the model outputs a score $s(y; x, \theta)$ for each class $y \in Y$. In order to compute the conditional probability of y in the output layer, we apply a softmax function over the output of the model. Therefore, the conditional distribution of the label over input x and model parameters θ is defined as:

$$p(y \mid x, \theta) = \frac{s(y; x, \theta)}{\sum_{y' \in Y} s(y'; x, \theta)}$$

The target of training method is to maximize the log-likelihood over the training set:

$$\theta = \underset{\theta}{argmax} \sum_{x \in X} log\, p(y \mid x, \theta)$$

We use categorical cross entropy loss as the loss function of our model, which is defined as:

$$L_i = -\sum_{y' \in Y} t_{il} \, log(p_{il})$$

where t_{il} is one hot representation of the true label of training instance X_i, and p_{il} is the condition probability of the label l given a training instance X_i.

For our Knowledge-Aware Deep Dual Networks, adagrad [43] is used to optimize the loss function. The parameters are updated by:

$$\theta_t = \theta_{t-1} - \frac{\alpha}{\sqrt{\sum g_i^2 + \epsilon}} g_t$$

where α is the learning rate and ϵ is a smoothing parameter, g_t and θ_t are the gradient matrix and the parameters at the training step t. We initialize all the parameters with normal distribution.We set the batch size to be 200. To prevent overfitting, we further adopt a dropout [44] method with drop rate 50%.

VII. Evaluation

A. Dataset

In the experiment, we use two datasets extracted from the MIMIC III Clinical dataset [45], a freely accessible critical care database and it collects 2,083,108 medical reports from 46,520 ICU patients between 2001 and 2012. It is the only freely accessible critical database of ICU patients and it has spawned more than a decade with detailed information about the patients to help improve clinical research and medical data mining around the world.

While in-hospital mortality can be obtained in hospitals, the dataset contains detailed mortality information in Social Security Administration Death Master File [2] which we can use to get the out of hospital mortality information. Table I describes the detailed information about MIMIC III Clinical Database.

We use two different types of medical notes, one is from the test examination notes such as Radiology, Electrocardiography(ECG), and Echo. We ignore the sequential feature of notes, instead, we aggregate them together and construct our final dataset. The other is from nursing notes which act differently from the style of test examination. In order to distinguish the two datasets, we denote the dataset from test examination records as RAD dataset and the dataset from nursing records as NURSING dataset.

TABLE I: MIMIC III Description

Category	No.
Patients	46,520
Radiology	522,279
Echo	45,794
ECG	209,051
Nursing	223,556

B. Preprocessing

1) Text processing: In our research, we extract patients' notes during patients' last visit and patient's mortality outcomes are computed based on the patient's last discharge time and the dead time recorded in the dataset. Similar to [7], we exclude patients whose ages are under 18. In addition, we only use the notes prior to the patient's last discharge and exclude notes that are recorded after the death time. Vocabularies for each medical document are generated by first tokenizing the free text and lemmatizing the words in the texts and then removing stop words. For vocabulary dictionary generation and text tokenizing, we use keras text proprecessing tools[3] and we use onix stop word dictionary[4] to filter stop words.

2) Text Conceptualization: For UMLS concept extraction, we use MetaMap described in III-B. Different from previous work, we extract concepts straightly from the raw texts without tokenizing and removing stop words. This is because some stop words may be contained in UMLS concepts so that removing stop words may introduce more noises when extracting the concepts from the document. We use a python wrapper[5] for MetaMap and use its default settings to extract UMLS concepts.

As shown in Figure 1, the concepts extracted from medical notes may contain information not related to the medical domain. For example, the concepts in the second table on the left part of Figure 1 may contain little important information about a patient and these concepts can be filtered by their semantic types. Therefore, we filter the concepts that are not strongly related to the medical domain by using the semantic types of the concepts.

We extract both UMLS concepts and their positions in a raw text while a confidence score and a semantic type from UMLS are assigned with each UMLS concept at each raw medical note. The semantic types assigned to each UMLS concept extracted by Metamap are used to filter concepts whose types are not related to the medical domain. After that, we sort concepts upon its positions. To handle concepts with different positions, we assign a unique 2-tuple which contains each concept CUI and its corresponding position. Finally, we apply merging sort algorithm on the 2-tuple and get a concept list in which the concept is sorted by its position in raw medical notes. The detailed process of concepts extracted from the notes is illustrated in Figure 6.

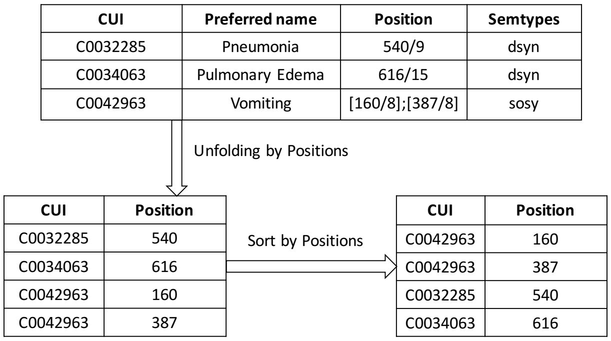
After the process of text preprocessing, text conceptualization, and removing the patients of whom the number of concepts is zero, 6,622 patients are left in the NURSING dataset and 35,263 patients are left in the RAD dataset. The detailed distribution of labels in the two datasets are illustrated in Table II and the number of words and concepts per document is illustrated in Table III, IV.

[2] https://www.ssdmf.com/

[3] https://keras.io/

[4] http://www.lextek.com/manuals/onix/stopwords.html

[5] https://github.com/AnthonyMRios/pymetamap



CUI	Preferred name	Position	Semtypes
C0032285	Pneumonia	540/9	dsyn
C0034063	Pulmonary Edema	616/15	dsyn
C0042963	Vomiting	[160/8];[387/8]	sosy

CUI	Position
C0032285	540
C0034063	616
C0042963	160
C0042963	387

CUI	Position
C0042963	160
C0042963	387
C0032285	540
C0034063	616

Fig. 6: Example of concept preprocessing: the concept indexed by C0042963 has two positions in the note, so we first unfold the concept and generate a tuple which contains concept CUI and its position and then sort the CUI by the position.

TABLE II: patient description on NURSING and RAD.

Hospital mortality	In hospital		Within 30 days		Within a year	
	pos	neg	pos	neg	pos	neg
NURSING	751	5,871	1,033	5,589	1,737	4,885
RAD	4,249	31,014	5,550	29,713	8,787	26,476

C. Experiment Settings

To illustrate the power of our Knowledge-Aware Deep Dual Networks, we randomly split the dataset with a ratio of 7:3 as the training set and test set. In order to find the best parameters of the model, we use 10% of the training set as the validation set.

TABLE III: dataset NURSING description.

No.	Mean	Std
Words per patient	160.25	101.91
Concepts per patient	51.13	31.18

TABLE IV: dataset RAD description.

No.	Mean	Std
Words per patient	1428.54	1700.138
Concepts per patient	170.658	134.9956

We use 50 filters in each convolutional neural network. As the number of patients is quite different in two datasets, in order to find the best of performances of our models, we use 100 as the embedding size of both the words and concepts on the dataset RAD and 20 on the dataset NURSING.

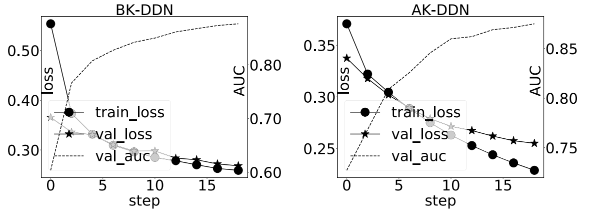
Our models are trained on a GPU server with one Nvidia Titan X GPU (pascal 12 G) and 2 CPU cores (2.1 GHz) in Ubuntu 14.04 platform. We implement our models and baselines based on keras which uses tensorflow as the backend.

For performance measurements, we report area under ROC curve (AUC) as our metric. We find that even the model is trained on an unbalanced dataset, the performance can also be competitive.

D. Baselines

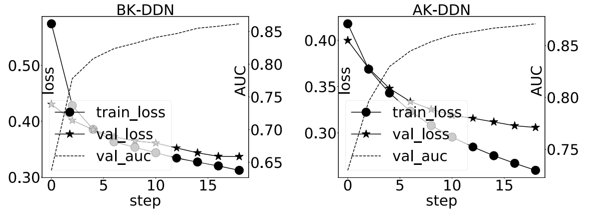
We compare our models with several state-of-the-art approaches: five feature-based methods and four deep learning based methods. The followings are the description of the baselines used.

- *LDA based word SVM.* The baseline uses the same parameters as [7]. Latent Topic Model(LDA) [12] is a topic model that models topic distributions of documents which can be used as a document representation. Therefore, we first train an LDA model with 50 topics to generate topic distributions of the document. And then we use the topic distributions as the document feature to train a Support Vector Machine (SVM) with a polynomial kernel. We use the LDA implementation in gensim and SVM implementation in sklearn.
- *LDA based concept SVM.* The baseline uses the same parameters (an LDA model with 50 topics) as *LDA based word SVM* to generate a concept-level topic distribution of a text. And then, the concept-level topic distribution of the document is used to train a Support Vector Machine.
- *LDA based word LR.* The LDA model is trained using the same parameters as *LDA based word SVM* to generate word-level topic distributions of a text. Then a Logistic Regression Model with L2 regularization is trained using the topic distributions learned from the LDA model.
- *Combined LDA with SVM.* We combine the concepts and the medical notes together as the dataset. And then use LDA to get the joint topic distributions to train a kernel support vector machine.
- *BoW + SVM.* We fit the support vector machine with the frequency of words in the dataset. In order to use most important words in one text, we use term frequency-inverse document frequency(tf-idf) metric to compute a score for each word in the dictionary and select top k most important words and count the frequency of the important words to construct a fixed-length of vector. And then, the support vector machine(SVM) is trained on the generated dataset. In our experiment settings, we set k to be 1000.
- *Text CNN.* Inspired by the method proposed in [15], we use the upper component of our BK-DNN with three convolutional layers and three max pooling layers. We concatenate the outputs of Max Pooling Layers and use a shape transformation matrix to mapping the concatenation to a vector of size 2.
- *Concept CNN.* The baseline uses the same parameters as Text CNN but uses concepts from the UMLS as its input. We use the lower part of the BK-DDN with the same parameters to generate a concept-level patient representation, and then a hidden layer is to map the concept-level presentation into a vector of size 2.
- *H_CNN.* The baseline is proposed in [11]. The baseline is a hierarchical embedding method in document classification using sentence level embedding to construct document level embedding and then use document level



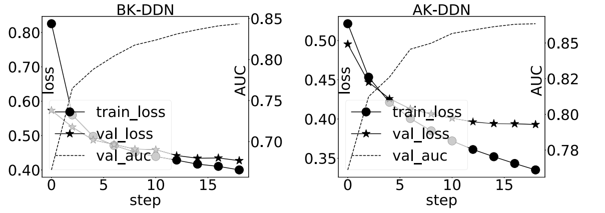
(a) BK-DDN (b) AK-DDN

Fig. 7: loss and AUC of validation set in in-hospital mortality prediction on RAD.



(a) BK-DDN (b) AK-DDN

Fig. 8: loss and AUC of validation set in hospital mortality within 30 days prediction on RAD.



(a) BK-DDN (b) AK-DDN

Fig. 9: loss and AUC of validation set in hospital mortality within year days prediction on RAD.

embedding for classification. As the source is not provided by the author, we implement the method ourselves.

- *DKGAM.* The baseline is proposed by Cao et al. [31]. In order to adapt it to our problem, we replace the words by the concepts and their positions. As the code is not provided by the author, we implement the model by ourselves.

E. Results

The results of the proposed AK-DDN and BK-DDN on two datasets are depicted in table V and table VI.

TABLE V: hospital mortality prediction on NURSING.

Models	$t = 0$	$t \leq 30$	$t \leq 365$
LDA based word SVM	0.756	0.738	0.721
LDA based word LR	0.811	0.788	0.738
BoW + SVM	0.815	0.797	0.766
LDA based concept SVM	0.756	0.690	0.669
Combined LDA with SVM	0.828	0.792	0.733
Text CNN	0.846	0.821	0.794
Concept CNN	0.825	0.785	0.796
H_CNN	0.802	0.772	0.751
DKGAM	0.811	0.790	0.775
BK-DNN	**0.848**	**0.821**	**0.805**
AK-DNN	**0.873**	**0.857**	**0.820**

TABLE VI: hospital mortality prediction on RAD.

Models	$t = 0$	$t \leq 30$	$t \leq 365$
LDA based word SVM	0.753	0.749	0.745
LDA based word LR	0.777	0.766	0.772
BoW + SVM	0.765	0.789	0.785
LDA based concept SVM	0.723	0.712	0.721
Combined LDA with SVM	0.802	0.782	0.774
Text CNN	0.847	0.851	0.824
Concept CNN	0.840	0.836	0.832
H_CNN	0.790	0.804	0.797
DKGAM	0.850	0.768	0.816
BK-DNN	**0.863**	**0.867**	**0.856**
AK-DNN	**0.880**	**0.873**	**0.862**

In the task of in hospital mortality prediction, the BK-DDN achieves an AUC of 84.8% on the NURSING dataset, which is competitive among the baselines. The AUC of BK-DDN is 86.3% on the RAD dataset which is above 1% higher than the best of the baselines (DKGRAM). Besides, the performance of AK-DDN is the highest, which achieves 87.3% on the NURSING dataset and 88% on the RAD dataset.

In the task of hospital mortality prediction within 30 days, the performances of the BK-DDN are 82.1% on the NURSING dataset and 86.7% on the RAD dataset, which shows its strengths among the baselines. What's more, the AK-DDN achieves the best of all the models and achieves 85.7% on the NURSING dataset (3% higher than the best of baselines) and 87.3% on the RAD dataset (2% higher than the best of baselines).

In the task of hospital mortality prediction with a year, the performances of BK-DDN are 80.5% on the NURSING (about 1% higher than that of Text CNN) and 85.6% on the RAD dataset (about 2% than that of the Concept CNN). In addition, the performances of AK-DDN are the best among the proposed models with an AUC of 82% on the NURSING dataset and 86.2% on the RAD dataset.

Compared the performances among Combined LDA with SVM, LDA based word SVM and LDA based concept SVM, the performance of combined LDA with SVM is higher than the other two baselines which suggests the combination of both words and concepts is promising and valuable. And the metrics(loss,AUC) during the training process of the three tasks are showed in Figure 7, 8, 9.

VIII. Discussion

To further explore the insights of our AK-DDN, we analyze the results of the RAD dataset. We show the power of our proposed model in the following ways: word and concept embedding analysis, patient embedding analysis.

A. Important Words and Concepts

In this section, we discuss word and concept embeddings trained in AK-DDN. Note that attention mechanism is utilized to generate both word-level text information and concept-level text information. Consider that in the generation part of concept based interaction generation in Section V-1 and word based interaction generation in Section V-2, each pair of the concepts and words per note is weighted by sorting the weights of the pairs. In order to further explore the power of attention mechanism, we simply take in hospital mortality prediction as our condition and select one positive sample (die in hospital) and one negative sample (not die in hospital) to visualize the weights of pairs.

Compared with Table VII and Table IX, consider that in Section V-2, we use each concept as a query to compute concept related word weights. From the above tables, we can dive into the insights of the attention mechanism. When it comes to the positive condition, the pairs are more related to disease and the status of disease(increased), this indicates that the patient's condition is getting worse. However, in the negative condition(not die in hospital), the pairs are more related with diseases and measurements(tube), this may indicate that some cure measurements are valid and the patient's condition is getting better.

Compared with Table VIII and Table X, the difference is that in the positive situation, the most important concepts and words are more related to diseases while in the other situation, those concepts act differently.

TABLE VII: important pairs in word based interaction (positive case).

Concept	Concept Definition	Word	Weight
C1527391	Anterior thoracic region	increased	0.1047
C0018802	Congestive heart failure	increased	0.0801
C0234438	Whiteout	pulmonary	0.0779
C0008031	Chest Pain	increased	0.0553
C0234438	Whiteout	ap	0.0534
C0549646	chest disorders	increased	0.0431
C0034063	Pulmonary Edema	increased	0.0431
C0234438	Whiteout	lung	0.0410
C0747635	Bilateral pleural effusion	increased	0.0308
C0013404	Dyspnea	increased	0.0248

TABLE VIII: important pairs in concept based interaction (positive case).

Concept	Concept Definition	Word	Weight
C0234438	Whiteout	pulmonary	0.1304
C0013404	Dyspnea	pulmonary	0.1296
C0234438	Whiteout	left	0.1231
C0242184	Hypoxia	pulmonary	0.0981
C0242184	Hypoxia	left	0.0927
C0013404	Dyspnea	pleural	0.0751
C0242184	Hypoxia	reason	0.0717
C0013404	Dyspnea	lung	0.0683
C0596790	interstitial	pulmonary	0.0650

TABLE IX: important pairs in word based interaction (negative case).

Concept	Concept Definition	Word	Weight
C0175730	biomedical tube device	tube	0.2304
C0596790	interstitial	tube	0.1538
C0242184	Hypoxia	tube	0.1119
C0185115	Extraction	chest	0.0891
C0336630	Endotracheal tube	tube	0.0888
C0015252	removal technique	chest	0.0882
C0013404	Dyspnea	tube	0.0869
C0332448	Infiltration	tube	0.0836
C0003873	Rheumatoid Arthritis	tube	0.0638
C0085678	Nasogastric tube	tube	0.0628

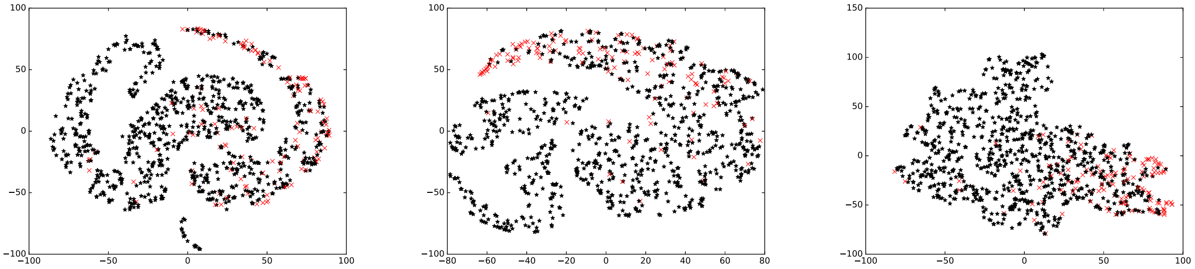
TABLE X: important pairs in concept based interaction (negative case).

Concept	Concept Definition	Word	Weight
C0015252	removal technique	enhancement	0.0723
C0015252	removal technique	axial	0.0656
C0728940	Excision	enhancement	0.0641
C0015252	removal technique	frontal	0.0587
C0728940	Excision	axial	0.0575
C0015252	removal technique	neck	0.0572
C0015252	removal technique	resection	0.0558
C0728940	Excision	frontal	0.0550
C0728940	Excision	resection	0.0499
C0015252	removal technique	post	0.0439

B. Patient Embedding

In this section, we analyze the patient embedding in the hidden layer of our AK-DDN. As illustrated in Section V, the final representation is composed of two parts: one is generated from word based interaction and the other from the concept based interaction. Therefore, for each patient in the dataset, three different types of information can be extracted from the output of the hidden layer in AK-DDN: word-level patient representation, concept-level representation, and combined patient representation. Then, we use T-SNE to visualize the first 1000 patients' representation.

Figures 10, 11, and 12 show the results of our AK-DDN in the three tasks: in-hospital mortality prediction, hospital mortality prediction within 30 days and hospital mortality prediction within a year. The left figure is the visualization of word-level patient representation, the middle is the visualization of concept-level patient representation and the right is the combined patient representation. Compared with the three pictures in Figures 10, 11, and 12, The major trends of the word-level and concept-level patient representations are aggregating in different directions which suggests both of the two representations model the patients in different semantic level. What's more, we can conclude from the pictures that points in the first two sub-figures do not show an action of clustering while patients' joint representations are similar in low dimension and tend to cluster in a direction. Therefore, the combination of concepts and text is useful to observe common information on the same label.

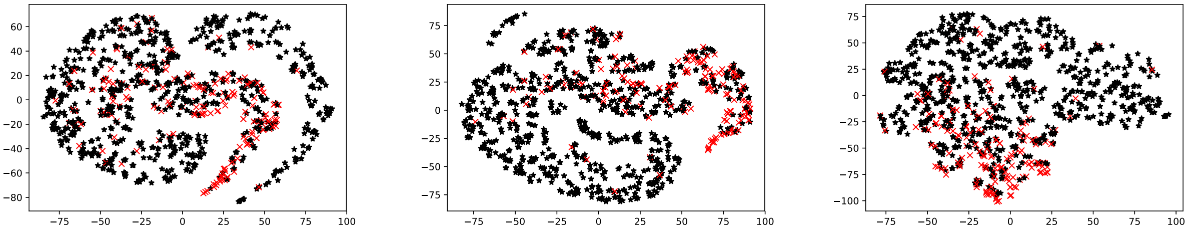


(a) word-level patient representation

(b) concept-level patient representation

(c) joint patient representation

Fig. 10: in hospital mortality prediction: Visualization of first 1000 patients' representation using T-SNE: Points marked with "×" and red color are positive labels and the others are negative labels.

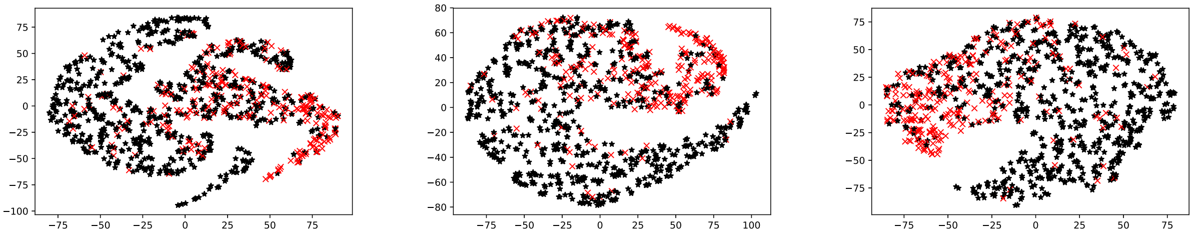


(a) word-level patient representation

(b) concept-level patient representation

(c) joint patient representation

Fig. 11: hospital mortality prediction within 30 days: Visualization of first 1000 patients' representation using T-SNE: Points marked with "×" and red color are positive labels and the others are negative labels.



(a) word-level patient representation

(b) concept-level patient representation

(c) joint patient representation

Fig. 12: hospital mortality prediction within a year: Visualization of first 1000 patients' representation using T-SNE: Points marked with "×" and red color are positive labels and the others are negative labels.

IX. Conclusion

In this paper, we address the text-based mortality prediction problem by fusing the feature construction stage and model training stage into an unified framework. We propose two novel Knowledge-aware Deep Dual Networks, which combine the word-level representation with the concept-level representation for prediction and further incorporate the co-attention mechanism to make the two representations learn from each other. Experimental results on the two real-word datasets show that the proposed models outperform the state-of-the-art approaches and verify the benefit of the co-attention mechanism.

Acknowledgment

We thank the anonymous reviewers for their valuable comments. This work was supported in part by National Natural Science Foundation of China under Grant No. 61532010 and 61521002, National Basic Research Program of China (973 Program) under Grant No. 2014CB340505, Shanghai Chenguang Program (16CG24), and Shanghai Sailing Program (17YF1404500).

References

[1] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in *KDD 2015*, 2015, pp. 855–864.

[2] A. E. W. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," in *MLHC 2017, Boston, Massachusetts, USA, 18-19 August 2017*, 2017, pp. 361–376.

[3] A. E. Johnson, A. A. Kramer, and G. D. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy," *Critical care medicine*, vol. 41, no. 7, pp. 1711–1718, 2013.

[4] Y. Luo, Y. Xin, R. Joshi, L. A. Celi, and P. Szolovits, "Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements," in *AAAI*, 2016, pp. 42–50.

[5] K. T. Islam, C. R. Shelton, J. I. Casse, and R. Wetzel, "Marked point process for severity of illness assessment," in *MLHC 2017*, 2017, pp. 255–270.

[6] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, "Variational recurrent adversarial deep domain adaptation," 2016.

[7] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *KDD*. ACM, 2014, pp. 75–84.

[8] M. Ghassemi, T. Naumann, R. Joshi, and A. Rumshisky, "Topic models for mortality modeling in intensive care units," in *ICML*, 2012, pp. 1–4.

[9] L.-w. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of icu patients using topic models inferred from unstructured progress notes," in *AMIA*, vol. 2012. American Medical Informatics Association, 2012, p. 505.

[10] Y. Jo, N. Loghmanpour, and C. P. Rosé, "Time series analysis of nursing notes for mortality prediction via a state transition topic model," in *CIKM 2015*, 2015, pp. 1171–1180.

[11] P. Grnarova, F. Schmidt, S. L. Hyland, and C. Eickhoff, "Neural document embeddings for intensive care patient mortality prediction," *arXiv preprint arXiv:1612.00467*, 2016.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[14] J. Wang, Z. Wang, D. Zhang, and J. Yan, "Combining knowledge with deep convolutional neural networks for short text classification," in *IJCAI*. AAAI Press, 2017, pp. 2915–2921.

[15] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[16] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of machine Learning research*, vol. 2, no. Dec, pp. 139–154, 2001.

[17] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *ACL*. Association for Computational Linguistics, 2012, pp. 90–94.

[18] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[19] M. Post and S. Bergsma, "Explicit and implicit syntactic features for text classification," in *ACL*, vol. 2, 2013, pp. 866–872.

[20] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *ACL*, 2016, pp. 1480–1489.

[21] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik, "Political ideology detection using recursive neural networks," in *ACL*, vol. 1, 2014, pp. 1113–1122.

[22] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano *et al.*, "The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[23] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (saps ii) based on a european/north american multicenter study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.

[24] J. Vicent, R. Moreno, J. Takala, S. Willats, A. De Mendonca *et al.*, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Med*, vol. 22, no. 5, pp. 707–710, 1996.

[25] S. Bhattacharya, V. Rajan, and H. Shrivastava, "Icu mortality prediction: A classification algorithm for imbalanced datasets." in *AAAI*, 2017, pp. 1288–1294.

[26] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[27] S. Saria, G. McElvain, A. K. Rajani, A. A. Penn, and D. L. Koller, "Combining structured and free-text data for automatic coding of patient outcomes," in *AMIA*, vol. 2010. American Medical Informatics Association, 2010, p. 712.

[28] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *NIPS*, 2015, pp. 919–927.

[29] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun, "Rubik: Knowledge guided tensor factorization and completion for health data analytics," in *KDD*. ACM, 2015, pp. 1265–1274.

[30] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun, "Leap: Learning to prescribe effective and safe treatment combinations for multimorbidity," in *KDD*. ACM, 2017, pp. 1315–1324.

[31] S. Cao, B. Qian, C. Yin, X. Li, J. Wei, Q. Zheng, and I. Davidson, "Knowledge guided short-text classification for healthcare applications," in *ICDM 2017*. IEEE, 2017, pp. 31–40.

[32] L. Wang, W. Zhang, X. He, and H. Zha, "Personalized prescription for comorbidity," in *DASFAA*. Springer, 2018, pp. 3–19.

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

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[35] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," *arXiv preprint arXiv:1607.04423*, 2016.

[36] W. Zhang, W. Wang, J. Wang, and H. Zha, "User-guided hierarchical attention network for multi-modal social image popularity prediction," in *WWW*, 2018, pp. 1277–1286.

[37] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[38] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *AMIA*. American Medical Informatics Association, 2001, p. 17.

[39] M. Simmons, A. Singhal, and Z. Lu, "Text mining for precision medicine: bringing structure to ehrs and biomedical literature to understand genes and health," in *Translational Biomedical Informatics*. Springer, 2016, pp. 139–166.

[40] G. Soğancıoğlu, H. Öztürk, and A. Özgür, "Biosses: a semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, 2017.

[41] J. D. Koola, S. E. Davis, O. Al-Nimri, S. K. Parr, D. Fabbri, B. A. Malin, S. B. Ho, and M. E. Matheny, "Development of an automated phenotyping algorithm for hepatorenal syndrome," *Journal of biomedical informatics*, vol. 80, pp. 87–95, 2018.

[42] E. Coiera, M. K. Choong, G. Tsafnat, P. Hibbert, and W. B. Runciman, "Linking quality indicators to clinical trials: an automated approach," *International Journal for Quality in Health Care*, vol. 29, no. 4, pp. 571–578, 2017.

[43] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, pp. 160 035 EP –, 05 2016. [Online]. Available: http://dx.doi.org/10.1038/sdata.2016.35