DCMN: Double Core Memory Network for Patient Outcome Prediction with Multimodal Data

Yujuan Feng†, Zhenxing Xu‡, Lin Gan§, Ning Chen†, Bin Yu§, Ting Chen† and Fei Wang‡
†Department of Computer Science and Technology, Tsinghua University, Beijing, China
‡Weill Cornell Medical College, Cornell University, NY, USA
§American Air Liquide, Newark, USA
†fyj15@mails.tsinghua.edu.cn, {ningchen, tingchen}@mail.tsinghua.edu.cn,
‡{zhx2005, few2001}@med.cornell.edu, §{lin.gan, bin.yu}@airliquide.com

***Abstract*—More and more healthcare data are becoming readily available nowadays. These data can help the healthcare professionals and patient themselves to better understand the patient status and potentially lead to improved care quality. However, the analysis of these data are challenging because they are large-scale and heterogeneous, high-dimensional and sparse, temporal but irregularly sampled. In this paper, we propose a method called Double Core Memory Networks (DCMN) to integrate information from different modalities of the longitudinal patient data and learn a joint patient representation effective for downstream analytical tasks such as risk prediction. DCMN is designed not only to disentangle the temporal and non-linear intra-modal dependencies for the data within each modality but also to capture the long-term inter-modal interactions. DCMN models are the end-to-end memory networks with two external memory cores where each modality of data is compressed and stored. Each memory core has an information-flow controller named query to interact with an external memory module. In addition, we incorporate a gating mechanism into basic DCMN model to perform dynamic regulation of memory interaction. DCMN models have multiple computational layers (hops) allowing data of different modalities interacting with each other recurrently along with a mechanism of alternating access of external memory for each memory core hop-by-hop. We evaluate DCMN models on two outcome prediction tasks, including a mortality prediction on the public Medical Information Mart for Intensive Care III (MIMIC-III) database and a cost prediction on the Hospital Quality Monitoring System (HQMS) dataset. Experimental results demonstrate that our DCMN models are more competitive over the baseline methods in the multimodal prediction setting.**

***Keywords*-double-core memory networks, multimodal patient data, outcome prediction**

I. Introduction

With the wide adoption of electronic health record (EHR) systems in hospitals worldwide, the explosion in the amount and abundance of biomedical data brings tremendous opportunities and challenges for healthcare research. The EHR system is implemented not only to archive patients information for administrative purpose but also facilitate secondary use of its data for clinical informatics. In hospital, the health status of a patient could be assessed comprehensively from different views, which makes the data large and heterogeneous. For example, there are streaming signals from bedside equipment for patients' real-time monitoring in critical care units, vital signs are captured secondly or minutely reflecting the status of the body's vital function, laboratory tests or image tests are recorded when clinician order them for diagnosis or screening, etc. The richness and granularities in such massive amount of multimodal of data enable us to build more accurate intelligent systems to inform clinical decisions. Deep learning, as a new wave of artificial intelligence technologies, has been applied in many areas including healthcare research [1], [2]. Its success and popularity is largely due to its great capacity of handling large-scale data in an end-to-end way. However, most of the previous works on deep learning in healthcare have only exploited a single data modality [3]–[6], which may not be enough for complicated conditions that need evidence synthesis from different information sources.

It is thus demanding for integrative analysis of multi-modal data in healthcare. One of the challenges for this task is that data collected from different sources are usually temporal and with variant timescale and sampling frequencies. A single-lead electrocardiogram (ECG) recording sampled at the frequency of 125HZ could have 37,500 values in 5 minutes, which is humongous compared to total observations of low-density vital signs or discrete events. Techniques like re-sampling of high-density inputs and filling sparse inputs with smoothed missing values have been explored in many research to overcome the problems of multi-resolution and sampling irregularities of multimodal data, [3], [7]. Secondly, the multimodal data streams are usually asynchronous where all features have exact time stamps and need to be temporally aligned [8], which is a challenging process. Thirdly, irrelevant features or noise exist in different modalities, which creates additional challenge for data integration.

In view of the above challenges, we aim to develop an algorithm that can effectively learn the temporal dependencies within each time-variant data modality and capture non-linear interaction across different data modalities. Our method is called *double core memory network* (DCMN), which is an end-to-end multi-layer deep learning framework with a double-core memory module. In each layer, we select the most relevant information from the memory cores in a soft manner with an attention mechanism and learn the combined representation for further inference. We test the DCMN model on two prediction tasks in healthcare. The first one is to predict the risk of mortality for patients stayed in intensive care units (ICU)

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00030

with integrated discrete longitudinal clinical events and multi-resolution continuous time monitoring data. Patient mortality is a primary outcome in ICU, and early prediction of mortality is an essential problem in critical care research. The second task is the cost prediction with two types of sequential diagnostic codes and surgery codes extracted from the front sheet of inpatient EHR records. Accurate assessment of cost consumed in the hospital plays a vital role in improving the efficiency of management and quality of healthcare. Our proposed DCMN models demonstrate competitive performance in both tasks.

II. Related Work

A. Representation of discrete clinical events

EHR records are often composed of many discrete clinical events which are coded into several classification systems for billing and administrative purpose. Some examples of code systems include such as the International Classification of Disease (ICD) for diagnosis, Current Procedural Terminology (CPT) for procedures, RxNorm for drugs and Logical Observation Identifiers Names and Codes (LOINC) for laboratory results. Conventionally the EHR records are usually aggregated into vector based representation for analysis [9]. To explore the temporality among the clinical events, Wang *et al.* proposed a matrix based representation for each patient's EHR with one dimension corresponding to time and the other dimension representing the events [10], [11]. Cheng *et al.* [4] further developed a deep learning approach based on convolutional neural network (CNN) to perform analysis on EHR matrices to learn better representations. Xiao *et al.* proposed an algorithm combining the idea of topic modeling and recurrent neural network (RNN) to derive interpretable EHR representations [12]. Min *et al.* also conducted a comparative study on different machine learning models, including both conventional models and deep learning model variants, on the task of prediction of the hospital readmission risk of patients with chronic obstructive pulmonary disease (COPD) [13].

B. Representation of high-density streaming data

In critical care, real-time monitoring ECG signal is an effective invasive tool to supervise the health condition of patients. And ECG signal is an informative indicator of the risk of mortality in ICU. ECG analysis usually contains four steps including preprocessing, feature extraction, feature representation and classification. Feature representation for ECG signal is key to building an accurate diagnosis system. Traditional representation methods mainly extract various types of ECG features including QRS, statistical, morphological, wavelet features [14] or perform heart rate variability analysis (HRV) [15] on the preprocessed ECG signal. Recently, deep learning techniques have been applied to large-scale ECG data analysis [16]. Rajpurkar *et al.* [17] proposed a 34-layer deep (convolutional neural network) CNN to extract ECG feature representation and yielded a cardiologist-level arrhythmia detection performance. Zihlmann *et al.* [18] design a deep CNN and convolutional recurrent neural network (CRNN) architectures to process arbitrary-length ECG signal for atrial fibrillation (AF) classification. Besides, there are some works found that transformation of 1-dimensional ECG signal into two-dimensional spectrum can improve the performance of ECG signal classification. And transfer learning is helpful to learn better representations for ECG signal [19].

C. Representation of multimodal data with deep learning

Deep learning has been increasingly applied to representation learning of multimodal data in healthcare. Some works extend the single-modal learning to the multimodal setting by using different information fusion strategies to combine feature vectors learned by separate neural networks. For example, Purushotham *et al.* [3] used feedforward network (FFN) to handle one modality of non-temporal data and GRU network for the other modality of temporal data and concatenated them to get the final patient representation. Feng *et al.* [20] proposed a multi-channel CNN model (MGCNN) to extract representations of two sequential medical codes separately and concatenated them to obtain a patient representation for outcomes prediction. Works described above are representatives of the simple fusion strategy with a shared representation layer to capture correlations among multimodal inputs, and we summarize them as the MultiModal Deep Learning model (MMDL). Recently, many complex fusion strategies have been proposed for multimodal representation learning. Chung *et al.* [21] applied a dual LSTM model to extract features of each modality and use attention mechanism for further information fusion. Le *et al.* [22] proposed a dual memory neural computer, which is based on the DNC model [23], to integrate two modalities of sequential data at the event level. Xu *et al.* [7] proposed RAIM, a CNN-LSTM model with guided multi-channel attention, to integrate continuous monitoring data and discrete clinical events. In their method, high density temporal physiological data from different channels need to be aligned into the same time step and then channel-wised attention was computed for information integration. It is constraint by the scope of synchronous setting and precise time information must be provided for time alignment.

D. End-to-End Memory Network Overview

Memory Networks (MemNN) [24] and NMTs [22] are the classes of neural memory networks with an external memory module. MemNN is an end-to-end recurrent memory neural network which has been widely and successfully applied in Question/Answer (QA) oriented tasks [25] and healthcare [26] recently.

MemNN takes an input set $\boldsymbol{x}_i, ..., \boldsymbol{x}_n$ to be stored in the memory module, and a query $\boldsymbol{q}$ to recursively interact with external memory to output an answer $\boldsymbol{a}$. In each single-layer, MemNN first converts the entire set of $\{\boldsymbol{x}_i\}$ into memory vectors $\{\boldsymbol{m}_i\}$ simply using an input memory embedding matrix A. The query $\boldsymbol{q}$ is also embedded into an internal state u. Then it computes a probability vector $\boldsymbol{p}$ over the inputs in the embedding space by a softmax:

$$p_i = Softmax\left(\boldsymbol{u}^T \boldsymbol{m}_i\right) \tag{1}$$

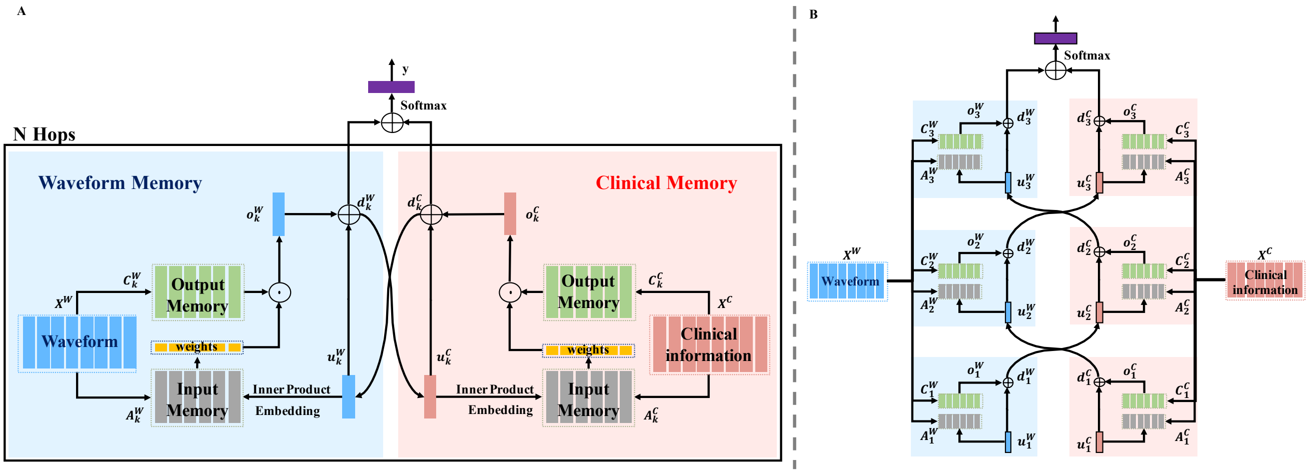

Fig. 1. A. Overview of Double-Core Memory Network. B. A three-layer version of DCMN model. For each patient with multimodal data such as waveform sequences and clinical sequences. Waveform data and clinical data extracted before the prediction time point are encoded and stored into two external memories of two memory cores. For each layer (hop), memory cores learn the interactions between the encoded query vector and external memory and produce combined outputs. In the multi-layer double-core memory setting, output information from two memory cores is used interactively. The output of waveform memory core is used as an input query to clinical memory core and the output of clinical memory core is used as an input query to waveform memory core for the following layer (hop). The final outputs of two memory cores at the last memory layer (hop) are combined to obtain a shared representation for the further prediction task.

where $Softmax(z_i) = e^{z_i / \sum_j e^{z_j}}$. A context vector $\boldsymbol{o}$ is a weighted sum of output vector $\boldsymbol{c}_i$ encoded from the memory core using an output memory embedding matrix C as:

$$\boldsymbol{o} = \sum_i p_i \boldsymbol{c}_i \tag{2}$$

The sum of the context vector $\boldsymbol{o}$ and the embedded query input $\boldsymbol{u}$ is the final output of memory core in a single-layer. In the multiple layers case, the output vector of layer k is fed to the layer $k+1$ as the query input of memory core, and the output of the last layer K is passed through the prediction layer to produce an answer:

$$\hat{\alpha} = Softmax(W(\boldsymbol{o} + \boldsymbol{u})) \tag{3}$$

III. Method

A. Problem Formulation

Firstly, let us formulate our multimodal learning problem $\{X_i^W, X_i^C, \boldsymbol{d}_i^S, y_i\}_{i=1}^N$, where N is the number of samples. Supposed each sample has two modalities of time-variant data $X^W = \{\boldsymbol{x}_t^W\}, t = 0, 1, ..., T_W$ and $X^C = \{\boldsymbol{x}_t^C\}, t = 0, 1, ..., T_C$, and one modality of time-invariant input variables $\boldsymbol{d}^S$, and an output y. More specifically, we focus on asynchronous setting where sequential data of two modalities are not required to be temporal alignment and can be different in time scale and length, that is $T_W \neq T_C$. Take one clinical setting for example, there are static demographic information $\boldsymbol{d}^S$, as well as two modalities of sequential inputs $X^W$and X^C, corresponding to the waveform data such as ECG signal and clinical time variant data such as vital signs, respectively.

B. Double Core Memory Network

We propose an end-to-end memory network called Double-Core Memory Network (DCMN) to deal with these multi-modal sequential inputs. As shown in Fig. 1A, DCMN is a multi-layer memory network with two external memory cores which allows two modalities of sequential data interacting with each other recursively.

1) Single Layer: We start by describing the DCMN model in the single-layer case. DCMN is composed of two symmetric memory cores: Waveform memory and Clinical memory. Each of memory core can be summarized into four modules: Input module, Attention module, Output module, and Prediction module.

Input module: Two modalities of inputs $X^W = \{\boldsymbol{x}_t^W\}$ and $X^C = \{\boldsymbol{x}_t^C\}$ are separately transformed into two series of memory vectors $M^W = \{\boldsymbol{m}_t^W\}$ and $M^C = \{\boldsymbol{m}_t^C\}$, which are then stored into two different external memory modules of Waveform memory core and Clinical memory core. Similar to memory core in the MemNN, we have one query vector $\boldsymbol{q}$ as the information controller to interact with external memory. More specifically, in the Waveform memory core, we compute the match between Waveform memory M^W and the internal state of query $\boldsymbol{u}^C$ encoded from clinical data X^C. Meanwhile, the internal state of query $\boldsymbol{u}^W$ embedded from Waveform data X^W is used to interact with Clinical memory M^C in the Clinical memory core. We can compute memory vectors via an input memory embedding matrix A in the way of $\boldsymbol{m}_t = A\boldsymbol{x}_t, t = 0, 1, ..., T$ or applying the LSTM network to encode temporal dependency in time-series inputs: $\boldsymbol{m}_t = \boldsymbol{h}_t = LSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}), t = 0, 1, ..., T$. The internal state of query $\boldsymbol{u}$ can also also embedded from the query vector $\boldsymbol{q}$ in a similar way.

Attention module: In this module, we apply an attentive mechanism to combine compressed information outputted from the external memory module. In each memory core, the match between the internal state of query $\boldsymbol{u}$ and each memory vector $\boldsymbol{m}_t$ is computed using (1). Then we compute a context vector $\boldsymbol{o}$ based on the attention weights $\boldsymbol{p}$ as in (2). The context vector is an attentive summarizing of output memory vectors calculated by $\boldsymbol{c}_t = C\boldsymbol{m}_t$, where C is an output embedding matrix.

Output module: In the single-layer case, the output $\boldsymbol{d}$ of each memory core is the combination of the internal state of query $\boldsymbol{u}$ and the context vector $\boldsymbol{o}$. This shortcut connection structure [27] alleviates gradient vanishing problem.

$$\boldsymbol{d} = \boldsymbol{u} + \boldsymbol{o} \tag{4}$$

Prediction module: Given output vectors $\boldsymbol{d}_K^W$ of Waveform memory core, $\boldsymbol{d}_K^C$ of Clinical memory core and the encoded static vector d^S, we further use the softmax to predict the label. The K is the number of layers of DCMN and $K = 1$ corresponds to the single-layer case.

$$\hat{y} = Softmax(U^W \boldsymbol{d}_K^W + U^C \boldsymbol{d}_K^C + \boldsymbol{w}^T d^S + b) \tag{5}$$

2) Multiple Layers: In the multi-layer case, w.r.t. $K > 1$, DCMN is a multi-layer architecture by stacking a number of double-core memory layers along with an alternating memory mechanism. Fig. 1B shows a three-layer version of our DCMN model. In this setting, each memory layer is named a hop.

- The output of a memory core at the layer k is computed as (6).

$$\boldsymbol{d}_k^W = \boldsymbol{u}_k^W + \boldsymbol{o}_k^W \tag{6}$$

$$\boldsymbol{d}_k^C = \boldsymbol{u}_k^C + \boldsymbol{o}_k^C \tag{7}$$

- Alternating external memory for memory cores hop-by-hop: The input to one memory core at the layer $k+1$ is the output of the other memory core at the layer k.

$$\boldsymbol{u}_{k+1}^W = \boldsymbol{d}_k^C \tag{8}$$

$$\boldsymbol{u}_{k+1}^C = \boldsymbol{d}_k^W \tag{9}$$

- Weight tying schemes: In the multi-layer case, the embedding matrices or the parameters can be shared across different hops, i.e. $A_1 = A_2 = ... = A_K$ and $C_1 = C_2 = ... = C_K$ for both Waveform and Clinical memory core. Optionally, each hop can have specific embedding matrices and be optimized independently.

3) Gated Double-Core Memory Network (Gated-DCMN): We explore two types of combination strategies to get output of memory core.

- The basic DCMN applies a linear mapping H to encoded internal state of query $\boldsymbol{u}$ before adding it to context vector $\boldsymbol{o}$:

$$\boldsymbol{d} = H\boldsymbol{u} + \boldsymbol{o} \tag{10}$$

- We proposed a variation of DCMN called Gated-DCMN, as shown in Fig. 2. It is designed to perform dynamic regulation of memory interaction in memory core, with the inspiration of the idea of the adaptive gating mechanism applied in GMemN2N [28]. The memory gate T is automatically learned based on the current internal state of query $\boldsymbol{u}$ and used to control the information flow in the memory core dynamically:

$$\boldsymbol{d} = T(\boldsymbol{u}) \odot \boldsymbol{u} + (1 - T(\boldsymbol{u})) \odot \boldsymbol{o} \tag{11}$$

$$T(\boldsymbol{u}) = \sigma(W\boldsymbol{u} + b) \tag{12}$$

where $\sigma(z)$ is sigmoid function and $\sigma(z) = 1/(1 + exp(z))$.

IV. EXPERIMENTS

A. Task 1: Mortality Prediction in ICU

We define a short-term mortality prediction task which is to predict whether the death happens within a short duration of time after the patient is admitted to the ICU. In our experiment, we extract a multimodal data from the data collection window, which is defined as the period of the first 24 hours after admitted to ICU, to predict whether the patient would die in the future 24 hours.

1) The MIMIC-III Database: We evaluate mortality prediction task on the Medical Information Mart for Intensive Care (MIMIC-III) database [1] which collects de-identified clinical data and physiological data from bedside patient monitors in adult and neonatal intensive care units (ICU) at the Beth Israel Deaconess Medical Center in Boston from 2001 to 2012 [29].

MIMIC-III Clinical Database: We extract one modality of clinical data from the MIMIC-III Clinical Database. It contains comprehensive information reflecting the state of a patient from multiple aspects, including demographics, vital sign measurements made at the bedside, laboratory test results, diagnoses, procedures, medications and patient outcomes (such as mortality and length of stay (LOS)).

MIMIC-III Waveform Database Matched Subset: The waveform database contains around 22,317 recordings of multiple physiologic signals (”waveforms”) and 22,247 time series of vital signs (”numerics”) and 10,282 matched and time-aligned clinical discrete records. The ’waveforms’ records are almost always composed of one or more multi-leads ECG signals, fingertip photoplethysmogram (PPG) signals, and up to 8 waveforms simultaneously [30]. And we extract one mortality of ECG signal for the mortality prediction task.

2) Cohort Selection: We extract a patient cohort for mortality prediction task in the guidance of criterion. Firstly, we identify all adult patients who are 15 years old or older at the time of ICU admission. Secondly, we only use the first admission of ICU and all other later admissions are dropped. Thirdly, we ensure the multimodal setting by only including patient with both clinical information and waveform information recorded during the data collection window. As a result, we get a matched cohort of 6177 ICU stays for mortality prediction. In statistics, around 1.6% of ICU stays

[1] https://archive.physionet.org/physiobank/database/mimic3wdb/

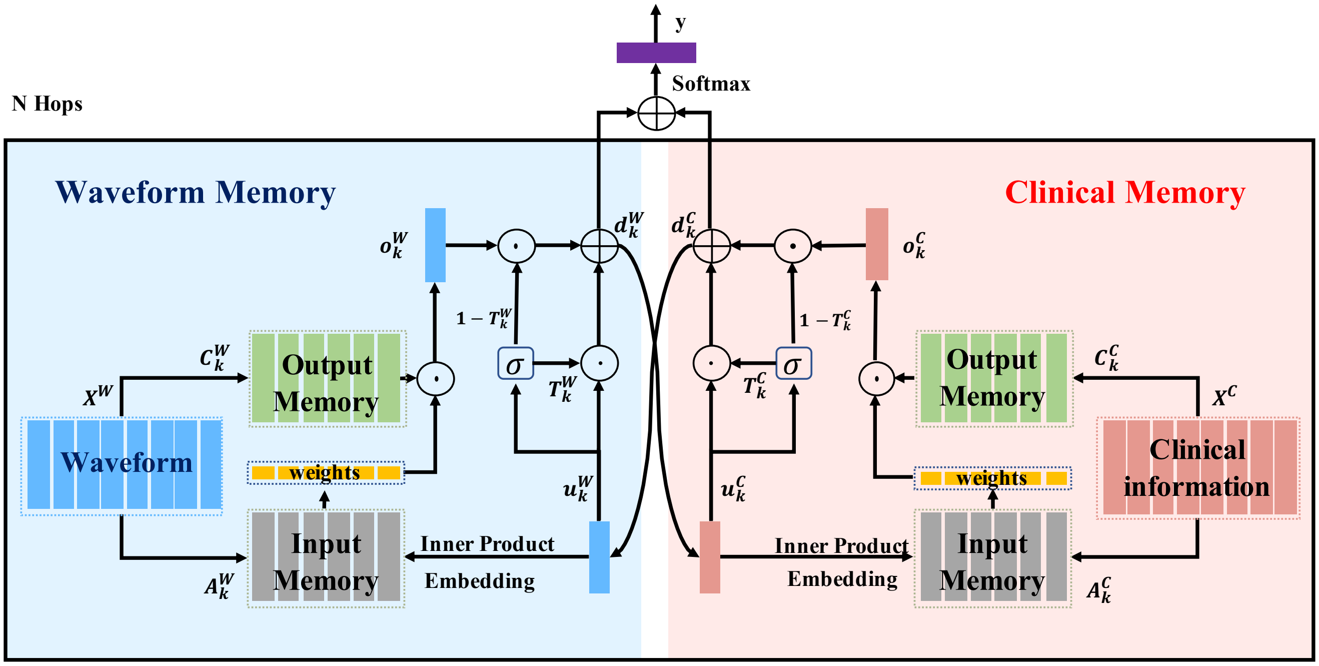

Fig. 2. The architecture of Gated Double-core Memory Network. In each memory core, there is a gate T to control the information flow dynamically. And the gate is based on current internal state of query $\boldsymbol{u}$.

has an outcome of death in the defined prediction window. The mortality prediction is a highly class imbalanced binary classification task.

3) Multimodal Data Preprocessing: In this section, we describe the steps of preprocessing and feature extractions for multimodal data.

Clinical data preprocessing: We use the clinical feature set A defined by the benchmark work done by Harutyunyan *et al* [5]. They performed a similar short-term mortality prediction task on the MIMIC-III database and achieved Area under the ROC curve (AUROC) score of 88.62%. The feature set A contains 17 important features including time-invariant features such as demographics and comorbidities, as well as time-variant features such as temperature and systolic blood pressure. All the variables are extracted from the data collection window and cleaned then preprocessed using the pipeline of benchmark work. Issues of outliers and inconsistent units, multiple recordings at the same time are addressed.

Clinical feature representation: We calculate five summary statistics for each time-variant clinical feature, including the minimum, maximum, average standard deviation, skewness and numbers of observations over several sub-windows. These sub-windows are sampled from the data collection window, covering the start, end, middle, whole part of data collection period. It handles the problem of irregular sampling to some extent. And all the categorical clinical features are one-hot encoded. These representation vectors of static and time-variant clinical features are concatenated and passed through baseline models to predict mortality.

For deep learning models, we represent time-variant features as a sequence of vectors. Given the $DCW = 24$ hours, each time-variant feature is re-sampled every 2 hours so that the time steps $T = 12$. During sampling, if there are multiple observations for a feature during the same 2-hour step, we use the last observation as the representative value. We fill-in missing value with forwarding imputation and use a mask to indicate whether the feature is missing or not. In this way, all selected time-variant features during each step is represented as a 76-dimensional vector. And we get a temporal representational matrix of shape (12, 76) for time-variant inputs. Min-max scalar normalization is used before we use LSTM architecture to encode clinical time-variant features into memory vectors for clinical memory core or get the internal state of query for waveform memory core in our DCMN models.

ECG signal preprocessing: For each ICU stay, We extract ECG lead-II signal during the data collection window from the waveform database. And the ECG signal is sampled at 125HZ. Raw ECG signal is normalized and then smoothed with a median filter to remove baseline wandering. FIR band-pass filter with cutoff frequencies 3-45HZ and Butter-worth low-pass filter with cutoff frequency 30HZ are applied to filter out motion artifact and noise.

ECG feature representation: ECG signal is representation of the periodic cardiac cycle. We perform heart rate variability (HRV) analysis on the preprocessed ECG signal. Heart rate variability describes complex beat-to-beat intervals variation and is a strong indicator of morbidity and death. We conduct three types of HRV analysis including time-domain methods, frequency-domain methods, and non-linear methods with the

NeuroKit Package [31]. Finally, we get a 39-dimensional ECG-level feature vector for each ECG sample. The HRV features are listed in Tabel I, and detailed descriptions can refer to the introduction of NeuroKit package [2].

Deep learning models have capacity to automatically learn robust representation from raw ECG signal, in ease of feature engineering. We first apply the short-time Fourier transform (STFT) to transform a ECG signal into a spectrogram. The spectrogram is the magnitude of the time-dependent Fourier transform versus time which conserves both temporal and frequency content of the signal. Then we use CRNN model described by Zihlmann *et al.* [18] to learn representation of ECG signal. The CRNN architecture combines convolutional layers for feature extraction with a LSTM layer for temporal aggregation of features. In our double-core memory network, the CRNN architecture is used to encode the ECG signal into sets of memory vectors for waveform memory core and an internal state of query for clinical memory core.

TABLE I
HANDCRAFTED HRV FEATURES FOR ECG SIGNAL

Type	Number	Feature Names
Time domain	10	RMSSD, meanNN, sdNN, cvNN,CVSD, medianNN, madNN, mcvNN, pNN50, pNN20.
Frequency domain	13	VLF, LF, HF, Triang, Shannon_h, ULF, VHF, Total_Power, LFn, HFn,LF/HF, LF/P, HF/P.
Non-linear	13	Entropy_SVD, DFA_1, DFA_2, Shannon,Sample_Entropy, Correlation_Dimension, Entropy_Multiscale_AUC, Entropy_Spectral_VLF, Entropy_Spectral-LF, Entropy_Spectral_HF, Fisher_Info, FD_Petrosian, FD_Higushi.
Others	3	n_Artifacts, ECG_Signal_Quality, Cardiac_Cycles_Signal_Quality.

B. Task 2: Cost Prediction in the hospital

We define a cost classification task which is to use two types of International Classification of Diseases (ICD) code sequences and additional demographic information collected from the electronic health records (EHRs) to predict the patient's total costs in the hospital. This cost prediction task is of great implication for the research of the Diagnosis-related Groups (DRGs) [32]. DRGs is a patient classification system where patients within each category are clinically similar and are expected to use the same level of hospital resources. And DRGs are classified based on diagnoses, procedures, age, gender, discharge status, and the presence of complications or comorbidities [33].

1) The Hospital Quality Monitoring System (HQMS): The HQMS system is the first official clinical data collection platform of China launched in 2011. It is originally designed to electronically collect the front sheet information in all inpatient medical records from tertiary hospitals every day. Each patient record usually consists of medical concepts (clinical diagnoses, surgical operations), demographic characteristics (age, gender) and outcome indicators such as total hospital costs and LOS. This digital information archived in the HQMS system can be secondarily used for healthcare applications.

2) Cohort Selection: We extract a set of tailored hospital records of nearly $800,000$ inpatients dated from Jan 2013 to Dec 2015 in the private HQMS database. A cohort of $597,396$ records is available after dropping incomplete records that without demographic characteristics or cost information. We filter out marginal records of which costs are outliers and obtain the final cohort of 550704 patient records for prediction. Variable of cost in the cohort has a mean of 18213.47 yuan with a standard deviation of 18777.56 yuan and a range of $[1406.36, 94735.72]$. Then records are categorized into four classes according to the value of cost. The classification system is defined with four ranges $[1406.36, 5000]$, $(5000, 10000]$, $(10000, 50000]$, $(50000, 94735.72]$ corresponding to four classes, respectively.

3) Multimodal Data Processing: Within an inpatient record, we have one main diagnosis (ICD-10 code) and, at most ten secondary diagnoses (ICD-10 code) with, at most ten procedure codes (ICD-9 code). The classification system used in the HQMS is the ICD-10 code (RC020 ICD-10, Beijing Version) for diagnosis and ICD-9 code (RC022 ICD-9, Beijing Version) for procedures. We can simply represent each ICD code as a one-hot vector. It is a binary vector of length of V with only one position being one to indicate the presence of code, and V is the size of the ICD codes corpus. The one-hot vector in the raw Specific level is high-dimensional and very sparse due to the size of ICD code corpus and existing of rare codes which seldom occur in the cohort. In the extracted cohort, there are 15,709 unique ICD codes including 11280 unique ICD-10 codes and 4426 unique ICD-9 codes in the Specific level. To overcome this problem, we map each raw ICD code in the Specific level into the Category level, which is a more generalized taxonomy in the ICD code system. The ICD code corpus size in the category level is reduced to 1469. In addition, we use the word2vec technique to embed each ICD code into a lower-dimensional space where semantically similar codes are also closed to each other. Given the embedding vector of each ICD-code, we can compute the representation of patient by concatenating or averaging embeddings of composing ICD codes. Optionally, we can first use the CNN architecture with pre-trained word2vec embeddings to learn representations of ICD-9 code sequence and ICD-10 code sequence separately and then add a shared layer to learn joint representation for cost prediction. In our DCMN model, the word2vec embeddings can be used to initialize the memory vectors and query vectors for two memory cores: diagnose memory core and procedure memory core, then be fine-tuned during optimizing the target loss.

[2] https://neurokit.readthedocs.io/en/latest/documentation.html

C. Prediction Algorithms

We compare our approaches [3] to several baselines.

- Logistic Regression with L1 Regularization (LR).
- Extreme Gradient Boosting (XGBoost).
- Multimodal Deep Learning Model (MMDL): The MMDL is proposed by Purushotham *et al.* [3] for benchmark mortality prediction on the MIMIC-III dataset. The main idea is to use different neural networks to get representations of each modality of data and then combined them with a fusion strategy such as concatenation or a shared representation layer. In mortality prediction task, we separately use CNN-LSTM model for ECG signal embedding and LSTM model for time-variant clinical sequences embedding. In cost prediction task, we apply two separate CNN architectures proposed by Feng *et al.* [20] to learn embeddings of two types of ICD code sequences. The shared representation layer is used for information fusion.
- End-to-End Memory Network (MemNN): We adapt the MemNN to learn representation of multimodal data. One modality of data can be encoded as a query and be used to recurrently interact with external memory. And the other modality of data is encoded and stored into external memory of memory core. After multiple rounds of the query, the output of memory core at the last layer is the representation of multimodal data. The neural networks used to encode each modality of data in MemNN model and DCMN models have the same setting as that in the MMDL model.

D. Experimental Setup

We implemented LR model with Scikit-learn 0.18.2 [34] and GBDT model with XGBoost software library [35]. The Word2vec model is implemented with the gensim package [36]. And embedding sizes for memory-based networks are set as 50 and 600 for mortality prediction and cost prediction, respectively. Deep models are implemented with keras 2.2.4 [37] and trained on workstations with NVIDIA GeForce GTX 1080 Ti. Stochastic Gradient Descent (SGD) optimizer with batch size 24, 1024 is used for mortality prediction task and cost prediction tasks, respectively. The numbers of hops for memory-based models range from 1 to 5. Best parameters are selected by grid search. Besides, we apply the learning rate decay schedule with the initial learning rate set to 0.001 and the early stopping strategy to accelerate the convergence. Batch normalization and dropout with a probability of 0.2 are also applied. Moreover, we use the up-sampling and compute the balanced class weights during training to overcome the class imbalance problem in the mortality prediction task. We use AUROC and Area under the Precision-Recall Curve (AU-PRC) with 5-fold cross-validation to report the results.

V. Results and Discussion

A. Task 1: Mortality Prediction in ICU

1) Quantitative Results: In this subsection, we describe the quantitative results of the short-term mortality prediction task.

Table II shows the macro AU-ROC score and macro AU-PRC score of different models for short-term mortality prediction using multimodal data. We observed that:

TABLE II
Mortality Prediction with Multimodal MIMIC-III Data

Algorithm	AU-ROC	AU-PRC
LR	0.8562	0.1522
XGB	0.8655	0.1603
MMDL [3]	0.8760	0.1738
MemNN [24]	0.8906	0.3045
DCMN	0.9038	0.3005
Gated-DCMN	**0.9195**	**0.3135**

- Deep learning models (e.g., MMDL and DCMN) obtain better results than traditional classification algorithms in term of AUC-ROC and AU-PRC metrics. One potential reason is that deep models with LSTM encoder can capture long temporal dependencies among the clinical time-variant data and ECG signals, which could be beneficial to the prediction of mortality.
- Compared to MMDL model using simply concatenation strategy, the memory-based networks including MemNN, DCMN and Gated-DCMN models yield better results. One potential reason is that recurrent architecture of memory cores can flexibly extract matching information among different modalities of data and learn better joint representation for mortality prediction.
- The DCMN models which integrate both clinical and ECG information through the architecture of double memory cores are better than single-core memory network. Because DCMN models take turns to store different information into memory rather than considering only one modality of data over multiple rounds of memory interaction. Especially, Gated-DCMN with a gated mechanism achieves the best performance, which indicates that dynamic control of information flow in the memory core can learn better inter-modality interactions.

2) Qualitative Results: In this subsection, we describe the qualitative results for the short-term mortality prediction task.

Compared to the result of the mortality prediction with only clinical data, the integration of clinical information and ECG signal improve the performance as shown in Fig. 3A. It suggests that utilizing complementary information contained in both clinical and ECG data is beneficial to the mortality prediction in ICU. After observing that the ECG signal contains important information related to mortality, we further explore the impact factors of ECG signal for the mortality prediction. The XGboost model is used to obtain the important scores of all of 39 HRV features. In Fig. 3B, we plot the top 10 important features based on their importance scores.

[3] Gated-DCMN source code is available at https://github.com/fengyujuan/Gated-DCMN

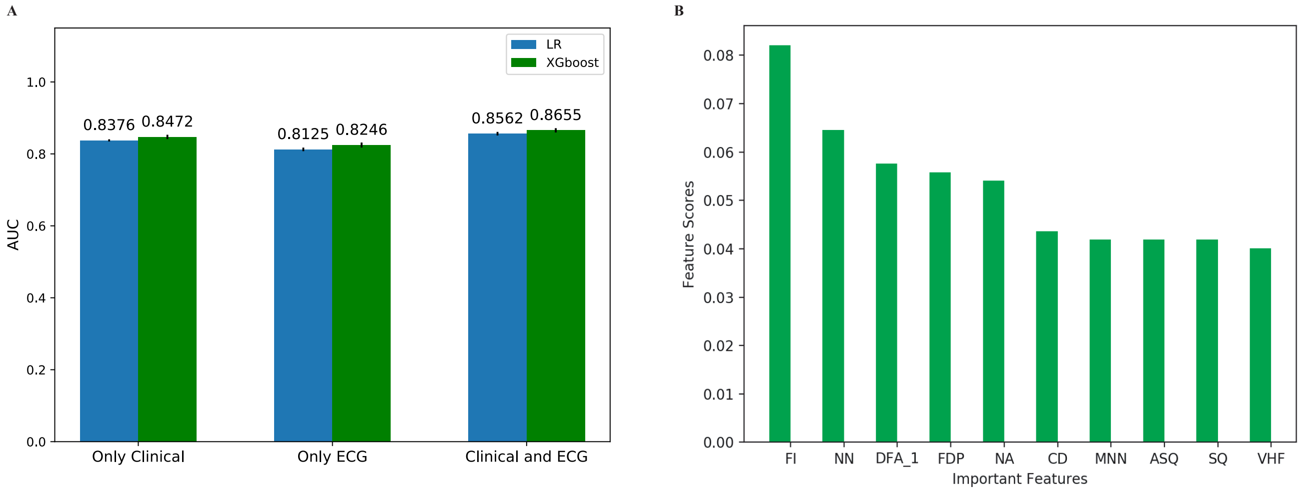

Fig. 3. A. Macro AUC for mortality prediction with different settings of data modality. B. The top 10 features selected from all HRV features based on importance score got by XGboost. FI: Fisher Information; NN: The mean RR interval. DFA: Detrended Fluctuation Analysis; DFA_1 is the short-term fractal scaling exponent calculated over n 4̄-16 beats. FDP: Petrosians Fractal Dimension over the RR intervals. NA: n_Artifacts; CD: Correlation Dimension; MNN: Median-based Coefficient of Variation; ASQ: Average Signal Quality; SQ: Signal Quality.

We can find that FI, NN, DFA and FDP are more important because they have a strong correlation with mortality. These results could corroborate well with some previous reports. For example, DFA was usually used as the predictor of mortality in patients [38], [39], which quantifies the presence or absence of fractal-like correlation properties [40] and provides some evidence in terms of the fluctuation of heartbeats [41].

B. Task 2: Cost Prediction in the Hospital

1) Quantitative Results: In this subsection, we summarized the quantitative results of the cost prediction task with the HQMS dataset.

TABLE III
Cost Prediction with the HQMS dataset

Method	Representation		Metric	
	Vector	*Dimension*	*AU-ROC*	*AU-PRC*
LR	one hot	1469	0.8554	0.6761
LR	word2vec	600	0.8468	0.6678
XGB	one hot	1469	0.8463	0.6859
XGB	word2vec	600	0.8833	0.7425
MMDL [20]	one hot	15706	0.8544	0.6757
MMDL [20]	word2vec	600	0.8839	0.7394
MemNN [24]	word2vec	600	0.8863	0.7546
DCMN	word2vec	600	0.8913	0.7568
Gated-DCMN	word2vec	600	**0.8932**	**0.7618**

From the results given in Table III, we can see that Gated-DCMN outperforms all other models in the task of cost classification, obtaining an average macro AU-ROC score of 89.32%, AU-PRC of 76.18%. What's more, the word2vec embedding method improves the performance over the one-hot encoding in different models. For ICD codes representation, the One-hot encoder learned from the Category level is better than that from the Specific level. This shows that incorporation of the domain knowledge into the representation system is particularly important.

We further test our models on a special cohort composed of rare testing patients which are defined as samples with more than one rare code. The rare code is defined as ICD code that occurs less than ten times in the HQMS dataset. The representations learning of rare patients with rare codes are much more difficult due to the data sparsity. We obtain average scores of macro AU-ROC of 82.22%, 82.53%, 83.86% and 84.42% and macro AU-PRC of 58.94%, 58.95%, 61.72% and 62.06% for the XGboost with word2vec embedding, the MMDL model, the MemNN models and our Gated-DCMN model, respectively. In Fig. 4, we plot the final patients' representations with t-SNE technique, it demonstrated that Our Gated-DCMN model can classify patients into more distinguishable groups while improving prediction performance.

We show an example of visualization of confusion matrix in Fig. 5 that using Gated-DCMN to predict cost and obtaining 90.16% and 77.83% for scores of AU-ROC and AU-PRC. It can be seen that it is easier to predict cost that ranges in $(10000, 50000]$ corresponding to class 2. As most of the patients stayed in the hospital consumed an average cost of around this range. And it is confused with extreme cases that patients cost little or too much in the hospital.

2) Qualitative Results: In this subsection, we compare the performance of the different number of hops K in our basic DCMN model for the cost prediction task, and we can get similar results of Gated-DCMN model. In the multi-layer case, we use a layer-wise weight tying scheme where the embedding matrices A and C in each of memory core are the same across different hops. In this setting, the DCMN models are RNN-like memory models which learn temporal inter-modality dependency recursively. We can observe in Fig. 6

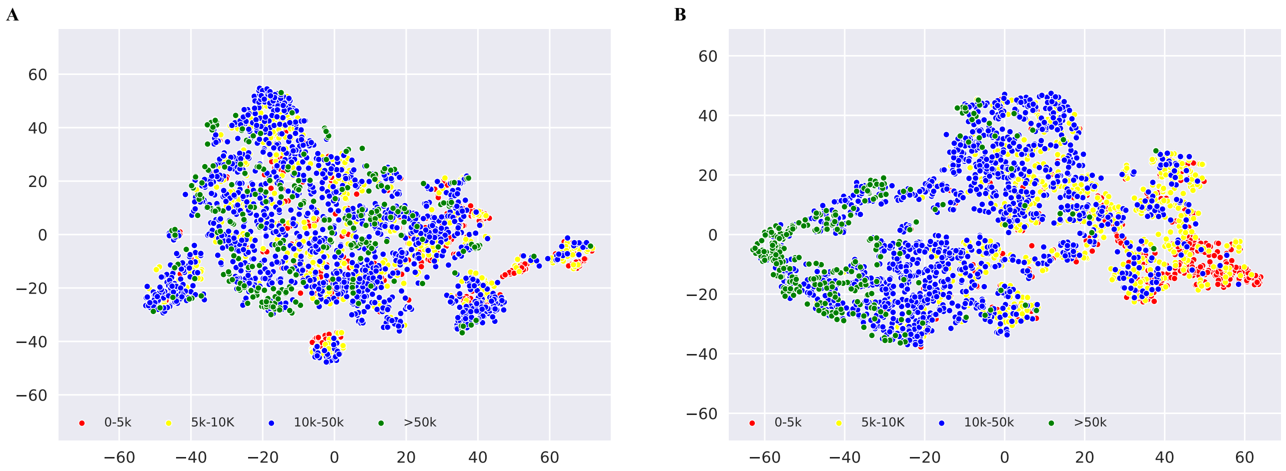

Fig. 4. The t-SNE visualization of the final patient's representations in cost prediction task. A: simply averaging of all codes' word2vec embeddings; B: the final joint representation learned by Gated-DCMN model. Class 0: 0-5k; Class 1: 5k-10k; Class 2:10k-50k and Class 3: >50k. And 1k means one thousand yuan.

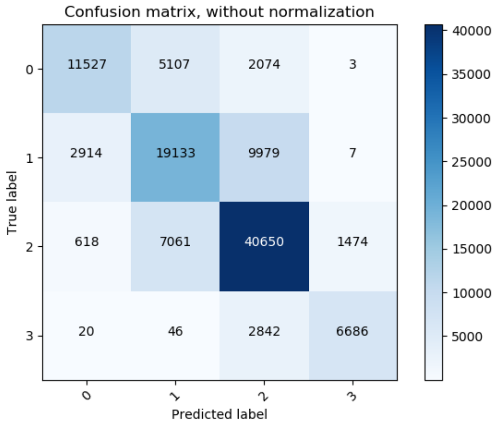

Fig. 5. Confusion matrix for cost prediction by Gated-DCMN model. Class 0: 0-5k; Class 1: 5k-10k; Class 2:10k-50k and Class 3: >50k. And 1k means one thousand yuan.

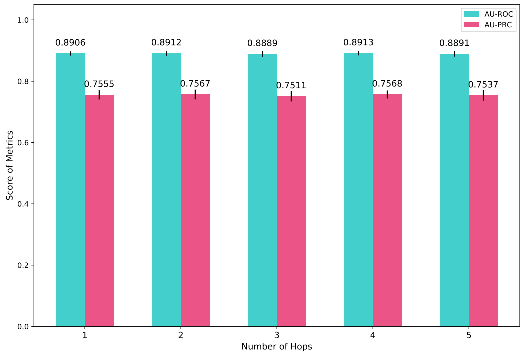

Fig. 6. Hyper Parameter tuning of the number of hops for basic DCMN model. Hops means layer.

that it tends to improve the performance as the number of hops increases although with little return. And the multi-layer interaction with the switching access of external memory facilitate the information fusion from different sources.

VI. CONCLUSIONS

Our proposed Double-Core Memory Networks are effective for learning from multimodal data sequences with different time resolutions. We evaluated the performance of our DCMN model on two different healthcare predictive modeling problems using two real world data sets, and our model is demonstrated to be able to outperform baseline models. In the future, a generalized multi-core memory network dealing with more data modalities (such as medical images or clinical notes) will be investigated for a more comprehensive assessment of health condition. Moreover, we will also develop effective interpretation method to explain why these models can work well.

ACKNOWLEDGMENT

The authors would like to acknowledge the support from Air Liquide Inc. on this research. The work of Yujuan Feng, Ning Chen and Ting Chen was also partially supported by the National Natural Science Foundation of China (61305066, 61561146396), the National Science Foundation of China (NSFC) of China Projects (Nos. 61673241, 61721003 and 61872218), and Tsinghua Fuzhou Institute research program and the Collaborative Project with Siemens.

REFERENCES

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.

[2] F. Wang, L. P. Casalino, and D. Khullar, “Deep learning in medicine-promise, progress, and challenges,” *JAMA internal medicine*, vol. 179, no. 3, pp. 293–294, 2019.

[3] S. Purushotham, C. Meng, Z. Che, and Y. Liu, “Benchmarking deep learning models on large healthcare datasets,” *Journal of biomedical informatics*, vol. 83, pp. 112–134, 2018.

[4] Y. Cheng, F. Wang, P. Zhang, and J. Hu, “Risk prediction with electronic health records: A deep learning approach,” in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 432–440.

[5] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, “Multitask learning and benchmarking with clinical time series data,” *arXiv preprint arXiv:1703.07771*, 2017.

[6] J. Xu, C. Deng, X. Gao, D. Shen, and H. Huang, “Predicting alzheimers disease cognitive assessment via robust low-rank structured sparse model,” in *IJCAI: proceedings of the conference*, vol. 2017. NIH Public Access, 2017, p. 3880.

[7] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, “Raim: Recurrent attentive and intensive model of multimodal patient monitoring data,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2565–2573.

[8] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, “Memory fusion network for multi-view sequential learning,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] J. Sun, J. Hu, D. Luo, M. Markatou, F. Wang, S. Edabollahi, S. E. Steinhubl, Z. Daar, and W. F. Stewart, “Combining knowledge and data driven insights for identifying risk factors using electronic health records,” in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 901.

[10] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. F. Laine, “A framework for mining signatures from event sequences and its applications in healthcare data,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 272–285, 2012.

[11] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, “Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach,” in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 453–461.

[12] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang, “Readmission prediction via deep contextual embedding of clinical concepts,” *PloS one*, vol. 13, no. 4, p. e0195024, 2018.

[13] X. Min, B. Yu, and F. Wang, “Predictive modeling of the hospital readmission risk from patients claims data using machine learning: A case study on copd,” *Scientific reports*, vol. 9, no. 1, p. 2362, 2019.

[14] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, “A survey on ecg analysis,” *Biomedical Signal Processing and Control*, vol. 43, pp. 216–235, 2018.

[15] S. N. Karmali, A. Sciusco, S. M. May, and G. L. Ackland, “Heart rate variability in critical care medicine: a systematic review,” *Intensive care medicine experimental*, vol. 5, no. 1, p. 33, 2017.

[16] P. Warrick and M. N. Homsi, “Cardiac arrhythmia detection from ecg combining convolutional and long short-term memory networks,” in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[17] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, “Cardiologist-level arrhythmia detection with convolutional neural networks,” *arXiv preprint arXiv:1707.01836*, 2017.

[18] M. Zihlmann, D. Perekrestenko, and M. Tschannen, “Convolutional recurrent neural networks for electrocardiogram classification,” in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[19] M. Salem, S. Taheri, and J.-S. Yuan, “Ecg arrhythmia classification using transfer learning from 2-dimensional deep cnn features,” in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2018, pp. 1–4.

[20] Y. Feng, X. Min, N. Chen, H. Chen, X. Xie, H. Wang, and T. Chen, “Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding,” in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 770–777.

[21] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[22] H. Le, T. Tran, and S. Venkatesh, “Dual memory neural computer for asynchronous two-view sequential learning,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1637–1645.

[23] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou *et al.*, “Hybrid computing using a neural network with dynamic external memory,” *Nature*, vol. 538, no. 7626, p. 471, 2016.

[24] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, “End-to-end memory networks,” in *Advances in neural information processing systems*, 2015, pp. 2440–2448.

[25] C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” in *International conference on machine learning*, 2016, pp. 2397–2406.

[26] X. Zhang, J. Chou, and F. Wang, “Integrative analysis of patient health records and neuroimages via memory-based graph convolutional network,” in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 767–776.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] F. Liu and J. Perez, “Gated end-to-end memory networks,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 1–10.

[29] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, p. 160035, 2016.

[30] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[31] D. Makowski, “Neurokit: A python toolbox for statistics and neurophysiological signal processing (eeg, eda, ecg, emg...).” 2016.

[32] R. B. Fetter, Y. Shin, J. L. Freeman, R. F. Averill, and J. D. Thompson, “Case mix definition by diagnosis-related groups,” *Medical care*, vol. 18, no. 2, pp. i–53, 1980.

[33] Wikipedia contributors, “Diagnosis-related group — Wikipedia, the free encyclopedia,” 2019, [Online; accessed 3-September-2019]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Diagnosis-related_group&oldid=911827409

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” pp. 785–794, 2016.

[36] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[37] F. Chollet, “keras,” https://github.com/fchollet/keras, 2015.

[38] T. H. Mäkikallio, S. Høiber, L. Køber, C. Torp-Pedersen, C.-K. Peng, A. L. Goldberger, H. V. Huikuri, T. Investigators *et al.*, “Fractal analysis of heart rate dynamics as a predictor of mortality in patients with depressed left ventricular function after acute myocardial infarction,” *The American journal of cardiology*, vol. 83, no. 6, pp. 836–839, 1999.

[39] N. Hotta, K. Otsuka, S. Murakami, G. Yamanaka, Y. Kubo, O. Matsuoka, T. Yamanaka, M. Shinagawa, S. Nunoda, Y. Nishimura *et al.*, “Fractal analysis of heart rate variability and mortality in elderly community-dwelling peoplelongitudinal investigation for the longevity and aging in hokkaido county (lilac) study,” *Biomedicine & pharmacotherapy*, vol. 59, pp. S45–S48, 2005.

[40] T. H. Mäkikallio, P. Barthel, R. Schneider, A. Bauer, J. M. Tapanainen, M. P. Tulppo, G. Schmidt, and H. V. Huikuri, “Prediction of sudden cardiac death after acute myocardial infarction: role of holter monitoring in the modern treatment era,” *European heart journal*, vol. 26, no. 8, pp. 762–769, 2005.

[41] H. V. Huikuri and P. K. Stein, “Heart rate variability in risk stratification of cardiac patients,” *Progress in cardiovascular diseases*, vol. 56, no. 2, pp. 153–159, 2013.