DMMAM: Deep Multi-source Multi-task Attention Model for Intensive Care Unit Diagnosis

Zhenkun Shi[1,2,3], Wanli Zuo[1,2], Weitong Chen[3], Lin Yue[1,4], Yuwei Hao[1,2], and Shining Liang[1,2(✉)]

[1] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
[2] College of Computer Science and Technology, Jilin University, Changchun, China
{shizk14,haoyw16,liangsn17}@mails.jlu.edu.cn, wanli@jlu.edu.cn
[3] The University of Queensland, Brisbane, QLD 4072, Australia
uqwche12@uq.edu.au
[4] Northeast Normal University, Changchun 130024, China
yuel031@nenu.edu.cn

Abstract. Disease diagnosis can provide crucial information for clinical decisions that influence the outcome in acute serious illness, and this is particularly in the intensive care unit (ICU). However, the central role of diagnosis in clinical practice is challenged by evidence that does not always benefit patients and that factors other than disease are important in determining patient outcome. To streamline the diagnostic process in daily routine and avoid misdiagnoses, in this paper, we proposed a deep multi-source multi-task attention model (DMMAM) for ICU disease diagnosis. DMMAM exploits multi-sources information from various types of complications, clinical measurements, and the medical treatments to support the diagnosis. We evaluate the proposed model with 50 diseases of 9 classifications on an extensive collection of real-world ICU Electronic Health Records (EHR) dataset with 151729 ICU admissions from 46520 patients. Experiments results demonstrate the effectiveness and the robustness of our model.

Keywords: Electronic Health Record · Disease prediction · Multi-source multi-task learning · Health care data mining

1 Introduction

The traditional model of clinical practice incorporates diagnosis, prognosis, and treatment. Diagnosis is fundamental to the practice of medicine and mastery of it is central to the process of both becoming and practicing as a doctor. Moreover, the activity of diagnosis is central to the practice of medicine, and has, to date, received the focused medical and computational science attention which many have argued it warrants [3]. This is beginning to be outburst with an emergent

G. Li et al. (Eds.): DASFAA 2019, LNCS 11447, pp. 53–69, 2019.
https://doi.org/10.1007/978-3-030-18579-4_4

computer-aided diagnosis, which seeks to explore the activity and its outcomes as a prism through which many issues are played out [14]. It is argued that diagnosis serves many functions for patients, clinicians, and wider society [14], and can be understood both as a category and a process [3]. Diagnosis classifies the sick patient as having or not having a particular disease. Historically, the diagnosis was regarded as the primary guide to treatment and prognosis ("what is likely to happen in the future"), and this is still considered the core component of clinical practice [8].

Intensive care refers to the specialized treatment given to patients who are acutely unwell and require critical medical care. Moreover, an **I**ntensive **C**are **U**nit (ICU) provides the critical care and life support for acutely ill and injured patients. The ICU is one of the most critically functioning operational environments in a hospital. To healing ICU patients, the clinicians need to actions in a remarkably short period. However, intensivists depend upon a large number of measurements to make daily decisions in the ICU. However, the reliability of these measures may be jeopardized by the effects of therapy [18]. Moreover, in critical illness, what is normal is not necessarily optimal. Diagnosis as the initial step of this medical practice is one of the most important parts of complicated clinical decision making [1].

With **E**lectronic **H**ealth **R**ecords (EHR) growth in biomedical and healthcare communities, it is possible to use bedside computer-aided diagnosis to accurate analysis of medical data, which can greatly benefit the ICU disease diagnosis as well as patient care, and community services. However, the existing work has focused on specialized predictive models that predict a limited set of disease. Such as Long *et al.* use the *IT2FLS* model to diagnosis heart disease [17], Jiri PolivkaJr *et al.* tried to find the mystery of the brain metastatic disease [22], Chaurasia *et al.* [4] use data mining techniques to detect breast cancer and Nilashi *et al.* [20] use neuro-fuzzy technique for hepatitis disease diagnosis. However, the day-to-day clinical practice involves an unscheduled and heterogeneous mix of scenarios and needs different prediction models in the hundreds to thousands [7]. It is impractical to develop and deploy specialized models one by one.

As shown in Fig. 1, this is the complication distribution of patients in the **M**edical **I**nformation Mart for **I**ntensive **C**are (MIMIC-III) [12]. We noticed that the vast majority of patients in the ICU are diagnosed with more than one diseases, that is to say, most of the patients have 5 to 20 complications. Moreover, the human body as organic entities and different systems are closely connected, and no diseases are isolated. In considering this, to establish a single model to diagnosis the majority of the diseases, we designed a multi-source multi-task attention [30] model for ICU diagnosis. The sources refer the different clinical measurements and the medical treatment, and the tasks refer the diagnose of different diseases, the detailed description will in the section of **Problem Definition**. To the best of our knowledge, this is the first time that to utilizing the shared feature space from different disease to boost the diagnose performance.

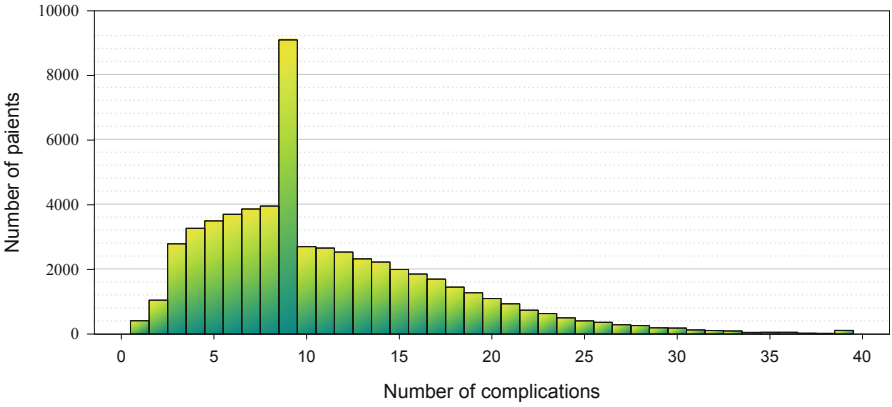

Fig. 1. Complication distribution of patients in MIMIC-III.

The focus of this paper is upon diagnosis as a process, we put the diagnosis into a temporal sequence and treated it as a step-by-step process, in particular from the perspective of the EHR data streaming. We conduct our experiment on real-world MIMIC-III benchmark dataset, and the result shows that our model is highly competitive and outperforms the state-of-the-art traditional methods and commonly used deep learning methods. Furthermore, we evaluated our model on 9 human systems over 50 different kinds of diseases.

The main contributions of this work are summarized as follow:

- **Multiple Perspectives for Disease Formulation.** We formulate ICU disease diagnose as a multi-source and multi-task learning problem, where sources correspond to clinical measurements and medical treatment, tasks correspond to the diagnosis of each disease. This work enables us to use a straightforward model to handle different kinds of diseases over all categories.
- **Diagnosis Step by Step.** For the first time, we treat the disease diagnosis as a gradual process over the observations along the temporal measure and treat sequence as well as the complications.
- **A Novel Integrated Model to diagnose the majority of the disease.** We designed a model DMMAM integrated with the input embedding, window alignment, attention mechanisms, and focal loss functions.
- **Comprehensive Evaluated Experiments.** We conduct experiment on MMIC-III benchmark dataset on 50 diseases over 9 categories, which covers most of the commonly diseases. The results demonstrate that our method is effective, competitive and can achieve state-of-the-art performance.

The remainder of this paper is organized as follows. We present a review of the recent advances in disease diagnoses briefly in Sect. 2. Section 3 gives out the detailed problem definition and our proposed framework. Section 4 introduced our experiment and our discussions. Section 5 concludes this study with future work.

2 Related Work

Diagnosis is the traditional basis for decision-making in clinical practice, inferring the disease from the observations attracts more and more attention in recent years [7,17,22,25,31]. Existing disease prediction methods can be roughly divided into two categories: clinical based diagnosis [9,22,25] and data based diagnosis [7,17,31]. Most existing clinical based diagnosis need profound knowledge of medical and most of them are focused on the certain field, such as specific diseases are caused by specific germs [21]. Until the last few years, most of the techniques for computer-aided disease diagnosis were based on traditional machine learning and statistical techniques such as logistic regression, support vector machines (SVM) [27], random forests (RF) [19] and decision tree (DT) [2,11,24]. Recently, deep learning techniques have achieved great success in many domains through deep hierarchical feature construction and capturing long-range dependencies in an effective manner [10]. Given the rise in popularity of deep learning approaches and the increasingly vast amount of clinical electronic data, there has also been an increase in the number of publications applying deep learning to diseases diagnosis tasks [5–7,20] which yield better performance than traditional methods and require less time-consuming preprocessing and feature engineering. For instance, Zhenping *et al.* [5] use the Best Mimic Model for ICU outcome prediction and got average 0.1 Area under Receiver Operating Characteristic (AUROC) score than SVM, LR and DT, Zachary C *et al.* learned to diagnose with long short-term memory (LSTM) recurrent neural networks and got average 0.5981 F1 scores over 6 different diseases.

However, all these methods are designed for a specific disease based on either the intensive use of domain-specific knowledge or taking advantage of advanced statistical methods. Specifically, studies have been conducted on Alzheimer's disease [31], heart disease [17], chronic kidney disease [28], and abdominal aortic aneurysm [13]. Moreover, these models have been developed to anticipate needs and focused on specialized predictive models that predict a limited set of diseases. However, the day-to-day clinical practice involves an unscheduled and heterogeneous mix of scenarios and needs different prediction models in the hundreds to thousands. It is impractical to develop and deploy specialized models one by one [7]. So it is significant to develop a unified model and can apply for the majority of diseases. This is beautiful dovetails to the multi-task learning, each disease can be treated as a single learning task. Note that many approaches to multi-task learning (ML) in the literature deal with a similar setting: They assume that all tasks are associated with the single output, e.g., the multi-class MNIST dataset is typically cast as 10 binary classification tasks. More recent approaches deal with a more realistic, heterogeneous setting where each task corresponds to a unique set of output [23]. We can not simply apply their approaches to ours, because we multiple clinical observations, multiple, and multiple medical treatments cannot be integrated into the existing frameworks.

More importantly, the human body as organic entities and different systems are intimately connected, and no diseases are isolated, so there may be little

difference between the complications. Therefore, based on our experiments it is hard for traditional methods to apply to such huge dataset over 50 kinds of diseases.

Inspired by the above problems, in this paper, we propose a general methodology, namely Deep Multi-source Multi-task Attention Model (DMMAM), to predict the disease from multi-modal data jointly. Here the sources indicate the clinical measurements and the medical treatments, the tasks represent the diagnosis of the diseases. In our work, the variables include not only the continuous clinical variables for regression (time series step by step regression) but also the categorical variable for classification (i.e., the class label for diseases classification). We treat the estimation of different diseases as different tasks, and multi-task learning [31] method developed in the machine learning community for joint learning. Multi-task learning can effectively increase the sample size that we are using to train the model because the samples of some kinds of disease are really small, which are not enough for learning (see Table 1). Specifically, at first, we assume that related tasks share a common relevant feature subset such as the age, temperature, heartbeat, blood pressure, *et al.* but with a varying amount of influence on each task, and thus adopt a hand engineered feature selection method to abstain a common feature subset for different tasks simultaneously. Then, we use a window alignment to adjust the time window between different sources and use one dense layer to reduce the dimensionality. Besides, we use two attention layer to capture the correlations between the different input sources as well as each time step. Finally, we use a gated recurrent unit (GRU) to fuse the above-selected features from each modality to estimate multiple regression and classification variables.

We will detail the problem definition in Sect. 3 and our proposed method in Sect. 4.

3 Proposed Framework

3.1 Problem Statement

For a given ICU stay length of T hours, and a collection of diagnostic results $R_t, t \in T$, it is assumed that we have a series of clinical observation:

$$O(t) = \begin{cases} R_t, & \text{if } R_t \notin \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $O(t)$ is vector of bedside observations at time t. $O(t) = P_a^i \Theta Q_b^i$, where P_a^i represent the i-th clinical measurement at time a, Q_b^j represent the j-th medical treatment at time b, and Θ is a window alignment operation between P_a^i and Q_b^j, and R_t represent the diagnostic result at time t. Our objective is to generate a sequence-level disease prediction at each sequence step. The type of prediction depends on the specific task and can be donated as a discrete scalar vector R_t^i for the multi-task classification. As all tasks are at least somewhat noisy, when

training a model $Task_i$, we expect to learn a good representation for $Task_i$ that ideally ignore the data-dependent noise and generalize well. By sharing representations between related tasks, we can enable our model to generalize better on our original task.

3.2 Multi-modal Multi-task Temporal Learning Framework for Temporal Data

Inspired by Daoqiang Zhang and Dinggang Shen's work [31], we treat the diagnosis of the diseases as a sequential multi-modal multi-task (SM3T) learning problem. The multi-modal represents the clinical measurements and the medical treatments. The tasks represent the diagnosis. The framework can simultaneously learn multiple tasks from multi-model temporal data. Figure 2 illustrates the proposed SM3T method and a comparison with the existing learning methods.

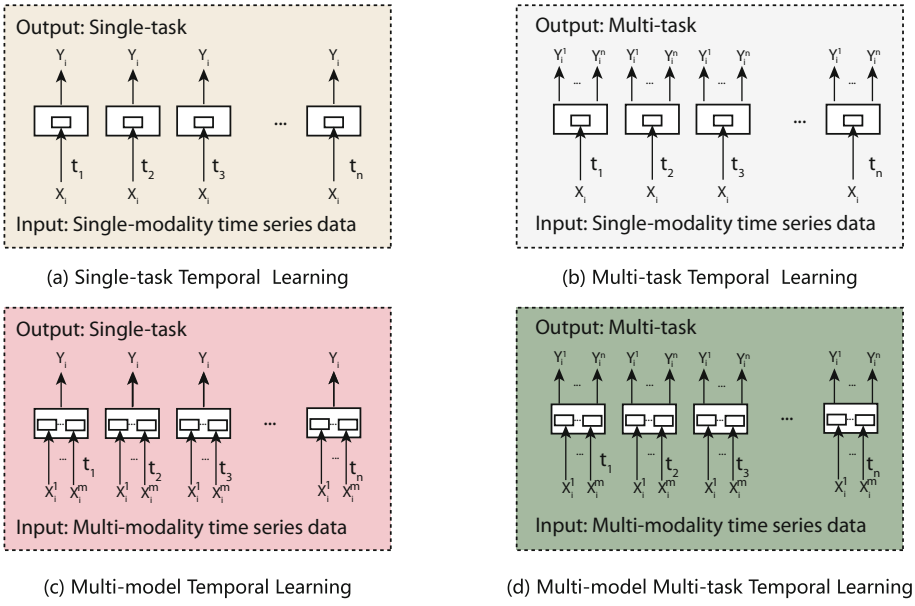

Fig. 2. Multi-modal multi-task temporal learning framework for temporal data.

Figure 2(a) is single-modality single-task temporal learning, each subject has only one modality of data represented as x_i at each time step, and each subject corresponds to only one task denoted as Y_i, this is the most commonly used learning method; Fig. 2(b) is single-modality multi-task temporal learning the input is similar as single-task temporal learning, but each object corresponds to multiple tasks denoted as $Y_i^1, Y_i^2, Y_i^3, ..., Y_i^n, n > 1$; Fig. 2(c) is multi-modality single-task temporal learning, each subject has multiple modalities of data represented as

$x_i^1, x_i^2, x_i^3, ..., x_i^n, n > 1$ at each time step and each subject corresponds to only one task denoted as Y_i; Fig. 2(d) is multi-modality multi-task temporal learning, each subject has multiple modality of data represented as $x_i^1, x_i^2, x_i^3, ..., x_i^n, n > 1$ at each time step and each subject corresponds to multiple tasks denoted as $Y_i^1, Y_i^2, Y_i^3, ..., Y_i^n, n > 1$.

Similar to Zhang's *et al.* [31] we can formally define the SM3T learning as below. Given N training subjects over T time span and each is having M modalities of data, represented as:

$$x_i^t = \{x_i^t(1), x_i^t(2), \ldots x_i^t(m), \ldots, x_i^t(M)\}, i = 1, 2, \ldots, N \tag{2}$$

our SM3T method jointly learns a series of models corresponding to Y different tasks denoted as:

$$Y_i = \{y_i^t(1), y_i^t(2), \ldots, y_i^t(j), \ldots, y_i^t(Y)\}, j = 1, 2, \ldots, N \tag{3}$$

Noting that SM3T is a general learning framework, and here we implement it through an attention framework as shown in Fig. 3. The x-axis represents the sequential data stream at time t, the y-axis represents the actions conducted on each t point and z-axis is the modalities of the input sources. In our experiment, $N = 2$ (e.g., $S1$ = clinical measurements and $S2$ = medical treatment) are used for jointly learning models corresponding to different tasks. We will detail the inner action of the SM3T framework in the following sections.

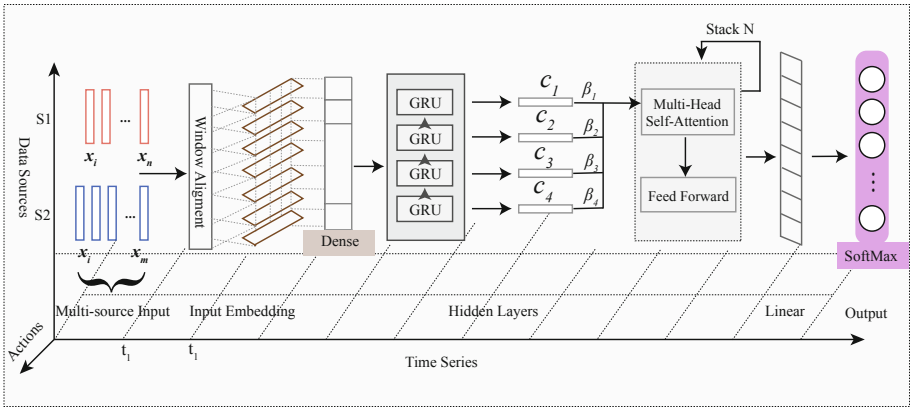

Fig. 3. The proposed multi-source multi-task attention model.

3.3 Input Embedding and Window Alignment

Give the R actions for each step for each step t, the first step in our model is to generate an embedding that captures the dependencies across different disease without the temporal information. In the embedding step, let N denote the number of diseases. The diagnosis process is first designed for each disease

without temporal information. Let P denote the ICU patients. The $p-th$ patient have h diagnosis results at time t, and $p-th$ patients with $h-th$ diseases is associated with two feature vectors $Sa_p^h(t)$ and $Sb_p^h(t)$ derived from the EHR, where $Sa_p^h(t)$ donate the clinical measurements and $Sb_p^h(t)$ donates the medical treatments. The dimension of Sa and Sb are α and β, respectively. Combined Sa and Sb, we generated a new feature vector Φ^h for the $p-th$ patient:

$$\Phi^p \equiv [\phi_1^p(t), \phi_2^p(t), \ldots, \phi_h^p(t)] \tag{4}$$

$$\phi_p^h(t) = \lambda_1^h Sa_p^h(t) \star \lambda_2^h Sb_p^h(t) \tag{5}$$

where $\star$ is **Window Alignment** operation, and λ_1 and λ_2 are trainable hyper-parameters for each disease.

Since our framework contains multiple actions, medical treatments Sb and clinical measurements Sa. The intentions of why we add a window alignment operation is that according to the common medical sense, the effect of treatment usually has some delay to the measurements. Assume $Sa_p^h(ti)$ represent the clinical measurements at time ti and $Sa_p^h(tj)$ represent the medical treatments at time step tj. The alignment is performed by mapping $Sa_p^h(ti)$ and $Sa_p^h(tj)$ into a unique time step $S_p^h(t)$. The alignment parameters λ_i^h are learned according to the patients and disease respectively. We found that tj usually later than ti, and this well accords with the prevailing medical sense.

3.4 Dense Layer

To balance the computational cost as well as the predictable performance, we need to reduce the dimensions before we transfer the raw medical data to the next process step. The typical way is to concatenate an embedding at every step in the sequence. However, due to the high-dimensional of the clinical features, "cursed" representation which is not suitable for learning and inference. Inspired by the Trask's work [29] in Natural Language Processing (NLP) and Song's [26] in clinical data processing, we add a dense layer to unify and flatten the input features. To prevent overfitting, we set dropout = 0.38 here.

3.5 The Gated Recurrent Unit Layer

The gated recurrent unit layer (GRU) takes the sequence of action $\{x_t\}_{t\geq 1}^T$ from the previous dense layer and then associate $p-th$ patient with a class label vector Y along with the time span, donates the class label for the $p-th$ patient with the $n-th$ disease at time T. $Y_p^n(t)$ is set ass follows:

$$Y_p^n(t) = \begin{cases} diseaseID, & \text{if diagnosis recorded at time } t \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

We create a T-dimensional response vector for the $p-th$ patient:

$$Y^{(p)} = (y_{p,1}, y_{p,2}, \ldots, y_{p,p_t})^\top \tag{7}$$

For the diagnosis of ICU patients, we adopted GRU and represent the posterior probability of the outcome of patient p has $y-th$ disease as:

$$Pr[P_y^n(t)=1|\phi_h^p(t)]=\sigma(\omega^{(p)^T}\phi_h^p(t)) \tag{8}$$

where $\phi(a)$ is the sigmoid function $\sigma(a)\equiv(1+\exp(-a))^{-1}$ and $\omega^{(p)}$ is a $\alpha+\beta$ dimensional model parameter vector for the $p-th$ patient.

To learn the mutual information of data resulting from the customization, we model for all disease jointly, so that we can share the same vector space across the disease, this is very useful for those diseases with fewer samples. We represent the trainable parameters of the GRU as a $(Sa+Sb)\times T$ $W\equiv[\omega^1,\omega^2,\cdots,\omega^t]$.

3.6 Multi-head Attention and Feed Forward

This attention layer is designed to capture the dependencies of the whole sequence, as we treated the diagnosis as a step-by-step process. In the ICU scenario, the actions (clinical measurements and medical treatments) closer to the current position are critical in helping the diagnosis. However, the observations further are less critical. Therefore, we should consider information entropy differently based on the positions which we make observations.

Inspired by [30], we use H-heads attention to create multiple attention graphs, and the resulting weighted representations are concatenated and linearly projected to obtain the final representation. Moreover, we also add 1D convolutional sub-layers with kernel size 2. Internally, we use two of these 1D convolutional sub-layers with ReLU (rectified linear unit) activation in between. Residue connections are used in these sub-layers. Unlike the previous work [1,4,7,11] making the diagnosis only once after a specific timestamp, we give out prediction at each timestamp. This is because the diagnosis results may change during the ICU stay and we make it as a dynamic procedure. This is more helpful for the ICU clinicians because they need to know the patients' possible disease at any time other than at the particular time. We stack the attention module N times and using the final representations in the final model. Moreover, this attention layer is task wise, that is to say if this attention will only work when this attention is helpful to the diagnosis.

3.7 Linear and Softmax Layers

The linear layer is designed to obtain the logits from the unified output of attention layer. The activation function used in this layer is ReLU. The last layer is preparing for the output based on different tasks. We use softmax to classify the different diseases, and the loss function is:

$$Loss\_d=\frac{1}{N}\sum_{n=1}^{N}-(y_k\bullet\log(\overline{y}_k)+(1-y_k)). \tag{9}$$

where N donate the number of diseases. Due to the distribution of the training set we also introduce Focal Loss as our loss function [16].

Table 1. Description of the prediction tasks based on ICD 9 code.

Category	ICD 9	Title	SampleSize	Age
1: Infectious and Parasitic	008.45	Int inf clstrdium dfcile	2672	69.07 ± 24.31
	038.9	Unspecified septicemia	5787	69.11 ± 32.13
2: Neoplasms	197.0	Secondary malig neo lung	866	62.23 ± 13.31
	197.7	Second malig neo liver	926	64.63 ± 17.47
	198.5	Secondary malig neo bone	984	63.59 ± 12.77
3: Endocrine, Nutritional, Metabolic and Immunity	250.00	DMII wo cmp nt st uncntr	10585	71.40 ± 28.41
	250.40	DMII renl nt st uncntrld	1574	69.26 ± 20.04
	250.60	DMII neuro nt st uncntrl	1793	70.02 ± 26.25
	263.9	Protein-cal malnutr NOS	2258	65.95 ± 26.35
4: Blood and Blood-forming Organs	280.0	Chr blood loss anemia	1346	68.34 ± 25.88
	280.9	Iron defic anemia NOS	1992	67.38 ± 39.21
	285.1	Ac posthemorrhag anemia	6998	69.10 ± 36.81
	285.21	Anemia in chr kidney dis	2616	66.70 ± 28.35
	285.29	Anemia-other chronic dis	2225	67.45 ± 32.21
	285.9	Anemia NOS	8253	67.90 ± 34.13
5: Circulatory System	397.0	Tricuspid valve disease	1286	77.26 ± 40.76
	401.9	Hypertension NOS	23153	71.27 ± 32.66
	403.90	Hy kid NOS w cr kid I-IV	4712	81.32 ± 45.61
	403.91	Hyp kid NOS w cr kid V	3756	65.27 ± 19.49
	410.71	Subendo infarct initial	4474	74.17 ± 30.51
	411.1	Intermed coronary synd	2200	69.42 ± 22.56
	412	Iron defic anemia NOS	4479	74.93 ± 36.99
	413.9	Angina pectoris NEC/NOS	1468	70.64 ± 27.84
	414.00	Crnry athrscl natve vssl	2415	78.53 ± 37.30
	414.01	Cor ath unsp vsl ntv/gft	14585	73.24 ± 32.09
	414.8	Chr ischemic hrt dis NEC	1526	74.54 ± 28.52
	431	Intracerebral hemorrhage	1561	69.71 ± 28.83
	433.10	Ocl crtd art wo infrct	1109	75.77 ± 30.39
	434.91	Crbl art ocl NOS w infrc	907	69.41 ± 28.22
6: Respiratory System	482.41	Meth sus pneum d/t Staph	1297	64.56 ± 22.81
	486	Pneumonia organism NOS	7779	68.51 ± 32.89
	491.21	Obs chr bronc w(ac) exac	1851	72.91 ± 24.79
	493.20	Asthma NOS	1215	69.22 ± 26.13
	493.90	Chronic obst asthma NOS	2781	59.18 ± 30.16
7: Digestive System	571.2	Cirrhosis of liver NOS	1529	55.93 ± 12.54
	571.5	Alcohol cirrhosis liver	1820	60.29 ± 16.73
8: Genitourinary System	584.5	Ac kidny fail tubr necr	3567	65.98 ± 24.11
	584.9	Acute kidney failure NOS	3564	71.45 ± 36.21
	585.6	Chronic kidney dis NOS	2720	62.39 ± 20.38
	585.9	End stage renal disease	4942	79.01 ± 41.90
	600.00	BPH w/o urinary obs/LUTS	1850	79.81 ± 35.58
9: Conditions originating in the perinatal period	765.18	Preterm NEC 2000-2499g	621	0.03 ± 0.03
	765.19	33-34 comp wks gestation	557	0.02 ± 0.02
	765.27	35-36 comp wks gestation	545	0.04 ± 0.03
	765.28	Preterm NEC 2500+g	642	0.02 ± 0.02
	769	Respiratory distress syn	511	0.10 ± 0.09
	770.6	Primary apnea of newborn	535	0.02 ± 0.03
	770.81	NB transitory tachypnea	331	0.10 ± 0.08
	774.2	Neonat jaund preterm del	1021	0.08 ± 0.08
	774.6	Fetal/neonatal jaund NOS	514	0.02 ± 0.04

4 Experiment

4.1 Data Description

We use a real-world dataset from MIMIC III[1] to evaluate our proposed approach. MIMIC-III is a large, publicly-available database comprising de-identified health-related data associated with approximately sixty thousand admissions of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The open nature of the data allows clinical studies to be reproduced and improved in ways that would not otherwise be possible [12]. In our experiment, we treat each ICU stay as a single case, because different ICU stay from the same patient may have diagnosed with a different disease. Moreover, this operation can help us to obtain more samples to train. As shown in Table 1, this is the first time that disease diagnosis conduct on such huge amount categories. We category the dataset based on the International Classification of Diseases (ICD) code, ICD-9, and we select 151729 ICU admissions over 50 commonly diagnosed disease. As shown in Fig. 1, most patients have multiple complications, and we collected all the complications in the whole ICU process temporally. Unlike the previous work, we did not filter any patients, this may results low performance, compared with related work. For the features, we included 529 clinical measurements features and 330 medical treatment features. Due to the abundant and mussy training samples, the performance between different disease is hugely different.

4.2 Experiment Settings

Our experiment includes over 40000 patients among 9 categories of 50 kinds of disease, the ICD9 code range from 001 to 779. A measure of the diagnosed disease, we set the outcome is "true" if the prediction result is right between the diagnose time span we observed the disease otherwise "false". In the training process, we will give out predict every time step only if there are observations during this time step, but in the test process we can give out diagnosis at every time step, and the time span can be customized. The learning rate in this experiment is 0.001, and the epochs size is 30. In our experiment, we set the batch size to 32, with ADAM optimizer and set dropout = 0.35. According to our experiment, we can get most of the best performance when then attention stack for 4 times. In order to conduct all the experiment in the same data, we manually divide the training set, validation set, and test set, we listed it in the Table 2.

4.3 Compared Methods

We compared our proposed method with 6 commonly used methods, i.e., Logistic Regression (LR) with L2 regularization, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), GRU, and the-state-of-the-art LSTM

[1] Data available at https://mimic.physionet.org/.

Table 2. Experiment settings for training, validating and test.

Task	Train	Validation	Test	Task	Train	Validation	Test
008.45	1870	534	268	414.8	1068	305	153
038.9	4050	1157	580	431	1092	312	157
197.0	606	173	87	433.10	776	221	112
197.7	648	185	93	434.91	634	181	92
198.5	688	196	100	482.41	907	259	131
250.00	7409	2117	1059	486	5445	1555	779
250.40	1101	314	159	491.21	1295	370	186
250.60	1255	358	180	493.20	850	243	122
263.9	1580	451	227	493.90	1946	556	279
280.0	942	269	135	571.2	1070	305	154
280.9	1394	398	200	571.5	1274	364	182
285.1	4898	1399	701	584.5	2496	713	358
285.21	1831	523	262	584.9	2494	712	358
285.29	1557	445	223	585.6	1904	544	272
285.9	5777	1650	826	585.9	3459	988	495
397.0	900	257	129	600.00	1295	370	185
401.9	16207	4630	2316	765.18	434	124	63
403.90	3298	942	472	765.19	389	111	57
403.91	2629	751	376	765.27	381	109	55
410.71	3131	894	449	765.28	449	128	65
411.1	1540	440	220	769	357	102	52
412	3135	895	449	770.6	374	107	54
413.9	1027	293	148	770.81	231	66	34
414.00	1690	483	242	774.2	714	204	103
414.01	10209	2917	1459	774.6	359	102	53

based method [15]. Due to the page limitation we only listed the two of the top two best methods in our paper. The first one is Logistic Regression (LR) with L2 regularization, and the second is the-state-of-the-art LSTM based method we listed LSTM+ in Table 3. As mentioned above, to ensure every evaluation method uses the same data, we fixed the dataset. As shown in Table 2 the validation and test date we use is approximately 25% of the whole dataset.

Table 3. Performance evaluation on each diagnose task.

Cat.	Task	LR			LSTM+			DMMAM(our method)		
		F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall
1	008.45	0.5822	0.6639	0.4784	0.8123	0.6840	**1**	**0.8641**	**0.7240**	**1**
	038.9	0.5822	0.6639	0.9345	0.7442	0.5259	**1**	**0.8641**	**0.8171**	0.9216
2	197.0	0.5593	0.5679	0.2414	0.7919	0.5392	0.7122	**0.8515**	**0.8281**	**0.8846**
	197.7	0.5895	0.5964	0.3333	0.7570	0.6357	0.8663	**0.8162**	**0.7172**	**0.8945**
	198.5	0.5366	0.5286	0.4500	0.6012	0.5214	0.5667	**0.6842**	**0.7118**	**0.6457**
3	250.00	0.5465	0.6546	**0.9754**	0.5443	0.6533	0.0531	**0.6545**	**0.7101**	0.6153
	250.40	0.8549	0.8941	0.0189	**0.9485**	**0.9021**	**1.0000**	**0.9485**	**0.9021**	**1.0000**
	250.60	0.8382	0.8892	0.0056	0.9413	0.8892	0.9000	**0.9613**	**0.9292**	**1.0000**
	263.9	0.8135	0.8602	0.0752	**0.9252**	**0.8608**	**1.0000**	**0.9252**	**0.8608**	**1.0000**
4	280.0	0.9139	0.9412	0.67454	0.6364	0.4116	0.5483	**0.9704**	**0.9425**	**1.0000**
	280.9	0.8740	0.9130	0.5050	0.9557	0.9152	0.9210	**0.9755**	**0.9347**	**1.0000**
	285.1	0.6405	0.6398	0.4037	0.8204	0.6995	0.9897	**0.8482**	**0.7412**	**0.9988**
	285.21	0.8531	0.8717	0.1832	0.9409	0.8883	0.8328	**0.9709**	**0.9283**	**1.0000**
	285.29	0.8589	0.9024	0.4327	**0.9503**	**0.9054**	0.9200	**0.9503**	**0.9054**	**1.0000**
	285.9	0.5216	0.5196	0.7167	0.5852	0.4996	0.5447	**0.7492**	**0.5533**	**0.8803**
5	397.0	0.8429	0.8587	0.1163	0.9453	0.8962	0.9991	**0.9457**	**0.8970**	**1.0000**
	401.9	0.5830	0.6232	0.1354	0.7277	0.6137	0.2380	**0.9213**	**0.9137**	**0.7612**
	403.90	0.6787	0.6621	0.6271	0.8065	0.6838	0.9027	**0.8255**	**0.7079**	**0.9466**
	403.91	0.7841	0.7892	0.4495	0.8777	0.7824	0.9956	**0.8795**	**0.7852**	**0.9993**
	410.71	0.7007	0.7062	0.3764	0.8293	0.7159	0.9314	**0.8333**	**0.7776**	**0.9499**
	411.1	0.7835	0.7670	0.5818	0.8559	0.7534	0.8887	**0.8705**	**0.7749**	**0.9177**
	412	0.6930	0.7131	0.2806	0.8122	0.6924	0.8951	**0.8382**	**0.7228**	**0.9668**
	413.9	0.8326	0.8611	0.1014	0.9345	0.8771	0.9937	**0.9376**	**0.8827**	**1.0000**
	414.00	0.6858	0.6536	0.5331	0.7307	0.6081	0.6588	**0.7419**	**0.6129**	**0.6894**
	414.01	0.5830	0.6232	0.2503	0.7606	0.6137	**1.0000**	**0.8508**	**0.7091**	**1.0000**
	414.8	0.8215	0.8468	0.1046	0.9327	0.8739	0.9955	**0.9350**	**0.8779**	**1.0000**
	431	0.8619	0.8540	0.5669	0.9317	0.8723	0.9954	**0.9332**	0.**8747**	**1.0000**
	433.10	0.8588	0.8899	0.0089	**0.9532**	**0.9106**	**1.0000**	**0.9532**	**0.9106**	**1.0000**
	434.91	0.8873	0.9058	0.0652	0.9123	0.9074	0.8975	**0.9619**	**0.9266**	**1.0000**
6	482.41	0.8705	0.9071	0.0153	0.9542	0.5091	0.8320	**0.9762**	**0.7210**	**0.9045**
	486	0.4328	0.5337	0.9345	0.6016	0.5210	0.0292	**0.9542**	**0.9125**	**1.0000**
	491.21	0.8180	0.8651	0.0269	**0.9338**	**0.8758**	**1.0000**	**0.9338**	**0.8758**	**1.0000**
	493.20	0.8782	0.9158	0.8861	0.9575	0.9185	0.8921	**0.9775**	**0.9432**	**1.0000**
	493.90	0.7508	0.7989	0.1039	0.8972	0.8136	0.8020	**0.9454**	**0.8732**	**1.0000**
7	571.2	0.5626	0.5685	0.4416	0.6866	0.5625	0.8846	**0.6951**	**0.5744**	**0.8956**
	571.5	0.5626	0.5685	0.6758	0.4111	0.5625	0.3377	**0.7204**	**0.6744**	**0.7455**
8	584.5	0.7662	0.7505	0.2817	0.9257	0.8618	**1.0000**	**0.9259**	**0.8620**	**1.0000**
	584.9	0.4235	0.5251	0.1003	0.5016	0.4945	0.0545	**0.7739**	**0.7173**	**1.0000**
	585.6	0.8768	0.8966	0.2169	0.9441	0.8941	1.0000	**0.9642**	**0.8943**	**1.0000**
	585.9	0.4858	0.4473	0.7859	0.8933	0.8072	**0.7892**	**0.9241**	**0.8124**	**0.8600**
	600.00	0.8984	0.9184	0.0919	**0.9637**	**0.9299**	**1.0000**	0.9627	0.9281	**1.0000**
9	765.18	0.8260	0.8555	0.0794	0.9361	0.8799	0.9979	**0.9372**	**0.8818**	**1.0000**
	765.19	0.8272	0.8593	0.5420	0.8446	0.8249	0.8213	**0.9357**	**0.8799**	**0.9769**
	765.27	0.8499	0.8518	0.2545	0.9446	0.8949	0.9979	**0.9456**	**0.8968**	**1.0000**
	765.28	0.8225	0.8255	0.2462	0.9340	0.8762	0.9979	**0.9359**	**0.8799**	**1.0000**
	769	0.8401	0.8349	0.2308	**0.9487**	**0.9024**	**1.0000**	**0.9487**	**0.9024**	**1.0000**
	770.6	0.8568	0.8518	0.3519	0.9476	0.9006	0.9192	**0.9486**	**0.9238**	**1.0000**
	770.81	0.9044	0.9343	0.9101	0.9680	0.9381	1.0000	**0.9691**	**0.9392**	**1.0000**
	774.2	0.6710	0.6379	0.4608	0.6174	0.4953	0.5035	**0.7904**	**0.6717**	**0.7657**
	774.6	0.8679	0.8593	0.4423	0.9497	0.9043	0.9102	**0.9765**	**0.9343**	**1.0000**

4.4 Evaluation Metric

To provide a comparison among the mentioned techniques, three evaluation techniques were used in this research: F1-Measure, Accuracy, and Recall. Those evaluation techniques are defined as:

$$\text{Accuracy} = \frac{TF + TN}{TP + FP + TN + FN} \quad \text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where TP and FP are the number of true positive and false negative, respectively.

4.5 Experiment Results and Discussions

Table 3 shows the prediction results. We can see that our model is significantly outperformed than all the baseline methods. Because we did not filter any ICU admissions and included all categories of the disease, so some evaluation metrics of our experiment are lower than those results appeared in Chen *et al.'s* work [15] (marked as LSTM+ in Table 3), but under the same experiment settings, our can always achieved the best performance. We can see that the number of the sample can greatly improve the diagnosis performance, the more samples, the better performance can achieve.

We discovered that the difference among categories are more evident than the diseases in the same category, and can pass average 3.2% in accuracy. The disease in category 3, Endocries, Nutritional, Metabolic and Immunity is the hardest disease to diagnosis in our model, and the disease of Conditions originating in the perinatal period in category 9 are the easiest ones to diagnosis. This is because there are greater diversities between category 9 and others, and there are smaller diversities between category 3 and others. Besides, the disease in the same categories have different diagnosis performance indicate that there is a higher relevance in the same system. We also conducted the ablation studies on the process of diagnosis, and the results show that the multi-source and multi-task can help us improved the performance among all the tasks over 3.6 percent in F1 scores. That is to say, by share the context feature space in the hidden layers the DMMAM can significantly improve the performance.

5 Conclusion and Future Work

In this study, we presented a new model named DMMAM for the disease diagnosis in the circumstances of the ICU. We modeled the ICU disease diagnosis as a multi-source multi-task classification problem. Moreover, we treat the diagnosis as a gradually process along the clinical measurements and the clinical treatments. The significances of our proposed model can be identified as:

1. **We considered the diversity of complications.** This both accords with the medical situation that no disease is isolated and different diseases have different diagnostic criteria and different treatment methods, the proposed multi-source multi-task model can perfectly suitable for this situations;
2. **We considered the diagnosis sequential relationship.** By introducing the attention layer we simulated the clinicians' diagnosis process and captured the interaction information among the sequence.
3. **Solved the imbalance problem.** The sample variance among the training data is hugely among different diseases. For example, the unspecified essential hypertension has 23153 samples. However, the secondary malignant neoplasm of the lung has only 866 samples. So if we are learning diagnosis without any precautionary measures, the diagnosis result would definitely to the majority ones. By using focal loss function, we alleviated problem caused by the unbalance of the dataset in the training process.

We conducted our experiment on 50 diseases over 167884 samples the results show the robustness and high accuracy. Moreover, this is the first time that diagnosis been conducted on such huge dataset. Nevertheless, how to use these diagnoses in further clinical actions remains a challenge in scientific research, and future work can be focused on this problem.

Acknowledgement. This work was supported by the Nature Science Foundation of Jilin Province (20180101330JC, 20190302029GX), the Fundamental Research Funds for the Central Universities (No. 2412017QD028), the China Postdoctoral Science Foundation (No. 2017M621192), the Scientific and Technological Development Program of Jilin Province (No. 20180520022JH, 20190302109GX). The authors also gratefully acknowledge the financial support from China Scholarship Council (No. 201706170617).

References

1. Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., Rashvand, P.: Diseases diagnosis using fuzzy logic methods: a systematic and meta-analysis review. Comput. Methods Programs Biomed. **161**, 145 (2018)
2. Azar, A.T., El-Metwally, S.M.: Decision tree classifiers for automated medical diagnosis. Neural Comput. Appl. **23**(7–8), 2387–2403 (2013)
3. Blaxter, M.: Diagnosis as category and process: the case of alcoholism. Soc. Sci. Med. Part A Med. Psychol. Med. Sociol. **12**, 9–17 (1978)
4. Chaurasia, V., Pal, S.: A novel approach for breast cancer detection using data mining techniques (2017)
5. Che, Z., Purushotham, S., Khemani, R., Liu, Y.: Interpretable deep models for ICU outcome prediction. In: AMIA Annual Symposium Proceedings, vol. 2016, p. 371. American Medical Informatics Association (2016)
6. Chen, M., Hao, Y., Hwang, K., Wang, L., Wang, L.: Disease prediction by machine learning over big data from healthcare communities. IEEE Access **5**, 8869–8879 (2017)
7. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor AI: predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference, pp. 301–318 (2016)

8. Del Mar, C., Doust, J., Glasziou, P.: Clinical thinking; evidence, communication and decision-making (2006)
9. Detemmerman, L., Olivier, S., Bours, V., Boemer, F.: Innovative PCR without dna extraction for African sickle cell disease diagnosis. Hematology **23**(3), 181–186 (2018)
10. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press, Cambridge (2016)
11. Hao, Y., Zuo, W., Shi, Z., Yue, L., Xue, S., He, F.: Prognosis of thyroid disease using MS-apriori improved decision tree. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) KSEM 2018. LNCS (LNAI), vol. 11061, pp. 452–460. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99365-2_40
12. Johnson, A.E., et al.: Mimic-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
13. Johnson, M.J., Willsky, A.S.: Bayesian nonparametric hidden semi-Markov models. J. Mach. Learn. Res. **14**, 673–701 (2013)
14. Jutel, A., Nettleton, S., et al.: Towards a sociology of diagnosis: reflections and opportunities. Soc. Sci. Med. **73**(6), 793–800 (2011)
15. Lin, C., et al.: Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 219–228. IEEE (2018)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
17. Long, N.C., Meesad, P., Unger, H.: A highly accurate firefly based algorithm for heart disease prediction. Expert Syst. Appl. **42**(21), 8221–8231 (2015)
18. Marshall, J.C.: Measurements in the intensive care unit: what do they mean? Crit. Care **7**(6), 415 (2003)
19. Nguyen, C., Wang, Y., Nguyen, H.N.: Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomed. Sci. Eng. **6**(05), 551 (2013)
20. Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., Akbari, E.: A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. J. Infect. Public Health **12**, 13 (2018)
21. Park, I.H., et al.: Disease-specific induced pluripotent stem cells. Cell **134**(5), 877–886 (2008)
22. Polivka, J., Kralickova, M., Kaiser, C., Kuhn, W., Golubnitschaja, O.: Mystery of the brain metastatic disease in breast cancer patients: improved patient stratification, disease prediction and targeted prevention on the horizon? EPMA J. **8**(2), 119–127 (2017)
23. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
24. Shi, Z., Zuo, W., Chen, W., Yue, L., Han, J., Feng, L.: User relation prediction based on matrix factorization and hybrid particle swarm optimization. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1335–1341. International World Wide Web Conferences Steering Committee (2017)
25. Sicherer, S.H., Sampson, H.A.: Food allergy: a review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. J. Allergy Clin. Immunol. **141**(1), 41–58 (2018)
26. Song, H., Rajan, D., Thiagarajan, J.J., Spanias, A.: Attend and diagnose: clinical time series analysis using attention models. arXiv preprint arXiv:1711.03905 (2017)

27. Subasi, A.: Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. Comput. Biol. Med. **43**(5), 576–586 (2013)
28. Tangri, N., et al.: A predictive model for progression of chronic kidney disease to kidney failure. JAMA **305**(15), 1553–1559 (2011)
29. Trask, A., Gilmore, D., Russell, M.: Modeling order in neural word embeddings at scale. arXiv preprint arXiv:1506.02338 (2015)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
31. Zhang, D., Shen, D., Initiative, A.D.N., et al.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage **59**(2), 895–907 (2012)