Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs

Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu and **Haifeng Huang**
Baidu Inc, Beijing, China
{chenjun22,daixiaoya,yuanquan02,luchao,huanghaifeng}@baidu.com

Abstract

The automatic text-based diagnosis remains a challenging task for clinical use because it requires appropriate balance between accuracy and interpretability. In this paper, we attempt to propose a solution by introducing a novel framework that stacks Bayesian Network Ensembles on top of Entity-Aware Convolutional Neural Networks (CNN) towards building an accurate yet interpretable diagnosis system. The proposed framework takes advantage of the high accuracy and generality of deep neural networks as well as the interpretability of Bayesian Networks, which is critical for AI-empowered healthcare. The evaluation conducted on the real Electronic Medical Record (EMR) documents from hospitals and annotated by professional doctors proves that, the proposed framework outperforms the previous automatic diagnosis methods in accuracy performance and the diagnosis explanation of the framework is reasonable.

1 Introduction

The automatic diagnosis of diseases has drawn the increasing attention from both research communities and industrial companies in the recent years due to the advancement of artificial intelligence (AI) (Liang et al., 2019; Esteva et al., 2019; Liu et al., 2018). As reported in (Anandan et al., 2019), "*AI-enabled analysis software is helping to guide doctors and other health-care workers through diagnostic processes and questioning to arrive at treatment decisions with greater speed and accuracy.*" Although the image-based diagnosis has been well studied using PACS (Picture Archiving and Communication Systems) data (Litjens et al., 2017), the text-based diagnosis for Clinical Decision Support (CDS) (Berner, 2007) remains difficult due to the rare access to reliable clinical corpus and the difficulty in balancing between accuracy and interpretability.

Table 1: A real outpatient EMR from hospital.

Section	Content
Basic	男, 30岁 (Male, 30 years old)
CC	咽部不适3天 (Pharyngeal discomfort for 3 days)
HPI	患者于3日前起咽痛伴发热, 无呼吸困难、咳嗽、咳痰、嗳气或反酸 (The patient developed pharyngalgia and fever 3 days ago, without dyspnea, cough, sputum, belching or acid reflux)
PE	咽峡稍充血, 双侧扁桃体Ⅰ度肿大, 无栓塞物及瘢痕 (The hypopharyngeal isthmus is slightly congested. The bilateral tonsils are first-degree enlarged. There is no embolism or scar in the pharynx.)
TR	血常规示白细胞计数升高, WBC12.5 $*$ 10^9/L. C反应蛋白正常. (The blood test showed elevated white blood cell count, WBC12.5 $*$ 10^9/L. The C-reactive protein is normal.)
Diagnosis	急性扁桃体炎 (Acute tonsillitis)

There have been attempts to study automatic text-based diagnosis with Electronic Medical Record (EMR) documents integrated in the Hospital Information System (Mullenbach et al., 2018; Yang et al., 2018; Girardi et al., 2018). Basically, an EMR document is written by a doctor and consists of several sections that describe the illness of the patient. Besides the patient's basic information like name, age and gender, an EMR document contains Chief Complaint (CC), History of Present Illness (HPI), Physical Examination (PE), Test Reports (TR, e.g. lab test reports and PACS reports), Diagnosis, etc. Table 1 shows a real outpatient EMR document from a hospital. These sections describe the patient's medical situation from different aspects: CC summarizes the patient's main discomforts of this visit. HPI extends CC by adding more details and findings from the conversation between doctor and patient. PE shows the findings by physically examining the patient's body, e.g. by palpation or inspection. TR are the objective findings from the lab test reports or the PACS reports. In the hospitals, the doctors will make a comprehensive analysis mainly based on CC, HPI, PE, TR and the basic information, and make a diagnosis. However, it is very hard for computers to automatically understand all the diverse sections and capture the key

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3143–3153
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

information before making an appropriate diagnosis. Besides, an *inpatient* EMR document is similar to that in Table 1 except that HPI, PE and TR are usually more lengthy and detailed. The framework proposed in this work can be applied on both the outpatient and the inpatient EMR documents and we will not distinguish them later.

In this study, we bring forward a novel framework of automatic diagnosis with EMR documents for CDS.[1] Specifically, we propose to predict the main diagnosis based on the patient's current illness. Different from the previous works (Yang et al., 2018; Sha and Wang, 2017; Li et al., 2017; Girardi et al., 2018; Mullenbach et al., 2018) that solely rely on the end-to-end neural models, we propose to stack the Bayesian Network (**BN**) ensembles on top of Entity-aware Convolutional Neural Networks (**ECNN**) in automatic diagnosis, where ECNN improves the accuracy of the prediction and BN ensembles *explain* the prediction. The proposed framework attempts to bring some interpretability of the predictions by incorporating the knowledge encoded in the BN ensembles. The main contributions of this work are as follows:

- We propose a novel framework that stacks the Bayesian network ensembles on top of the entity-aware convolutional neural networks to bring interpretability into automatic diagnosis without compromising the accuracy of deep learning. Interpretability is very important in the AI-empowered healthcare studies.
- We bring forward three variants of Bayesian Networks for disease inference that provides interpretability. Moreover, we ensemble these BNs towards more robust diagnosis results.
- The evaluation conducted on real EMR documents from hospitals proves that the proposed framework outperforms the previous automatic diagnosis methods with EMRs. The proposed framework has been used as a critical component in the clinical decision support system developed by Baidu, which assists physicians in diagnosis in over hundreds of primary healthcare facilities in China.
- We publish the Chinese medical knowledge graph of Gynaecology and Respiration used in our Bayesian Network for disease inference with this paper for reproducibility. The data set can be downloaded from Github.[2]

2 Related Work

Due to the rapid advancement of machine intelligence, the text-based automatic diagnosis is becoming one of the most important applications of machine learning and natural language processing in the recent years (Anandan et al., 2019; Koleck et al., 2019). Different from diagnosis or question answering on the Web (Chen et al., 2019), diagnosis for the CDS takes place in the hospitals and clinics, and the predictive algorithm is integrated into the Hospital Information System to assist doctors and physicians in the diagnosis.

Liang et al. (2019) proposes a top-down hierarchical classification method towards diagnosing pediatric diseases. From the root to the leaf, each level on the diagnostic hierarchy is a logistic regression model that performs classification on labels from coarse granularity to fine-grained granularity, e.g. from organ systems down to respiratory systems and to upper respiratory systems. This method requires heavy manual annotation of training samples at different levels of hierarchy.

Zhang et al. (2017) combines the variational auto-encoder and the variational recurrent neural network together to make diagnosis based on laboratory test data. However, laboratory test data are not the only resources considered in this paper.

Prakash et al. (2017) introduces the memory networks into diagnostic inference based on free text clinical records with external knowledge source from Wikipedia.

Sha and Wang (2017) proposes a hierarchical GRU-based neural network to predict the clinical outcomes based on the medical code sequences of the patient's previous visits. It deals with the sequential disease forecasting problem with EHR data rather than the diagnosis problem for the current visit with EMR document. Similarly, Choi et al. (2016a) studies the RNN-based model for clinical event prediction. Baumel et al. (2017) investigates the multi-label classification problem for discharge summaries of EHR with hierarchical attention-bidirectional GRU.

The most similar works to ours are in (Yang et al., 2018; Li et al., 2017) which trains an end-to-end convolutional network model to predict di-

[1] Different from Electronic Health Record (EHR) where the illness of a patient's multiple visits are combined together, EMR only contains the patient's illness of this particular visit. EMRs are more generally used in the hospitals in China.

[2] https://github.com/PaddlePaddle/Research/tree/master/KG/ACL2020_SignOrSymptom_Relationship

agnosis based on EMRs. Besides, Girardi et al. (2018) improves the CNN model with the attention mechanism in automatic diagnosis. Moreover, Mullenbach et al. (2018) studies a label-wise attention model to further improve the accuracy of diagnosis at the cost of more computation time. Choi et al. (2016b) proposes a reverse time attention mechanism for interpretable healthcare studies.

Different from the previous studies, the novelty of this paper is to bring interpretability into automatic diagnosis by stacking the ensembles of Bayesian networks on top of the entity-aware convolutional neural networks.

3 The Proposed Framework

Automatic diagnosis can be formally considered as a classification problem where the proposed method outputs a probability distribution $\Pr(d|\mathbf{S})$ over all diseases $d \in \mathbf{D}$ based on the illness description $\mathbf{S}$. In this study, $\mathbf{S}$ corresponds to the patient's EMR document, i.e. $\mathbf{S}$ consists of several sections of texts and some structured data like age, gender and medical department.

We bring forward a new framework that combines the black-box deep learning and the white-box knowledge inference to diagnose disease with EMR documents. Figure 1 shows the architecture of the proposed framework. Firstly, the medical entities are extracted from the EMR contents. Then, the EMR document is fed into the entity-aware convolutional networks to generate disease prior probability. Next, the Bayesian network ensembles perform disease inference based on the prior probability and the probabilistic graphical models (PGMs) before ensembling the final predictions.

3.1 Named Entity Recognition

Before introducing the convolutional and the Bayesian networks, we first discuss a basic component of this framework – the named entity recognition (NER). NER extracts the entities as well as their types from text sentences, which is very important to capture the key information of the texts. In our experiments, we used Baidu's enterprise Chinese medical NER system that integrates the advanced NER models (Dai et al., 2019; Jia et al., 2019) and extracts entities of *symptoms*, *vital signs*, *diseases* and *test report findings*.

The F1 score of the NER system we use is 91% in a separate evaluation conducted on 1000 deduplicated sentences from real EMR documents by 10 certificated physicians in China. [3] Meanwhile, the polarity (*positive (+)*, *negative (-)* or *unknown (?)*) of entities is also recognized. The polarity in this work objectively means the presence or absence of a finding in a given EMR. It is recognized in conjunction with the rule-based method with a vocabulary of negative Chinese words as well as the polarity detection model. Table 2 shows the NER results of the EMR in Table 1. Please note that the disease (*acute tonsillitis*) from the diagnosis section is the ground-truth label to predict and it will not be included in the input to the predictive model in the evaluation.

Table 2: The NER results of the EMR document shown in Table 1. TR Finding: test result finding. (+) for positive, (-) for negative and (?) for unknown.

Word	Section	Type	Polarity
咽部不适 (pharyngeal discomfort)	CC	Symptom	(+)
咽痛 (pharyngalgia)	HPI	Symptom	(+)
发热 (fever)	HPI	Symptom	(+)
呼吸困难 (dyspnea)	HPI	Symptom	(-)
咳嗽 (cough)	HPI	Symptom	(-)
咳痰 (sputum)	HPI	Symptom	(-)
嗳气 (belching)	HPI	Symptom	(-)
反酸 (acid reflux)	HPI	Symptom	(-)
咽峡充血 (congested hypopharyngeal isthmus)	PE	Vital Sign	(+)
双侧扁桃体肿大 (enlarged bilateral tonsils)	PE	Vital Sign	(+)
咽部栓塞物 (pharyngeal embolism)	PE	Vital Sign	(-)
咽部瘢痕 (pharyngeal scar)	PE	Vital Sign	(-)
白细胞计数升高 (elevated WBC)	TR	TR Finding	(+)
C反应蛋白异常(abnormal C-reactive protein)	TR	TR Finding	(-)
急性扁桃体炎 (acute tonsillitis)	Diagnosis	Diesease	(+)

In the offline processing of the EMR corpus, we preserved the Top-K most frequent entities of all types as the *entity vocabulary*. In later experiments, we empirically set $K = 10,000$. The entity vocabulary will be used to construct the one-hot feature for each EMR document, which will be introduced later. Since NER is not the focus of this study, the readers can choose the public Chinese NER API[4] from Baidu for fast experiments. We will focus on the major contributions of the proposed framework in the next sections.

[3] There are two senior physicians beyond the attending doctor level and eight junior physicians contributed in the annotation tasks here and later.

[4] http://ai.baidu.com/tech/cognitive/entity_annotation

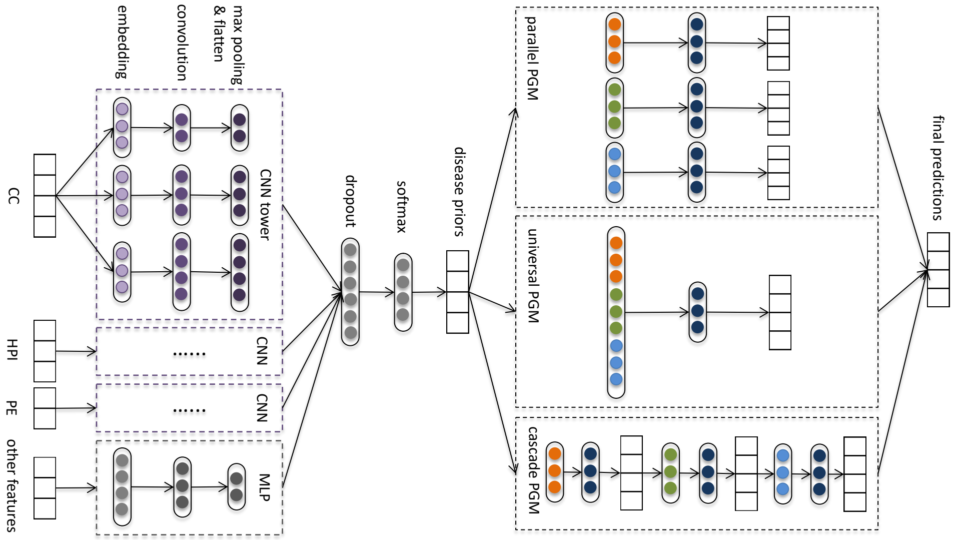


Figure 1: The architecture of the proposed framework.

3.2 ECNN for Prior Generation

The convolutional networks take as input the list of texts w.r.t. the sections of an EMR document as well as the medical entities extracted from them, and output the probability distribution of the diseases. To distinguish from the previous CNN models without medical entities (Yang et al., 2018; Li et al., 2017), we use **ECNN** to denote the entity-aware CNN model proposed in this paper where another branch of fully connected layers processes the medical entities and outputs the corresponding feature representation. Let N denote the number of sections (CC, HPI, PE, TR, etc) selected from the EMR document to construct ECNN. ECNN consist of two parts: (1) N convolutional towers, each of which reads a unique section, and (2) one multi-layer perceptron (MLP) branch that reads a high-dimensional hand-crafted feature.

Similar to the previous CNN method for text classification (Kim, 2014), each *convolutional tower* processes the input sequence with three kernels of various length resulting in multi-channel feature output. The three kernels process the input with 3-grams, 4-grams and 5-grams, respectively, and their outputs are concatenated as the output of a convolutional tower. Each kernel in the convolutional networks has 100 filters with strides as 1. The input is padded with *valid* method and the output is activated by ReLU.

For the input of MLP, we create the entity vocabulary that consists of the top-K frequent entities. Then, each EMR document is transformed to a K-dimensional one-hot feature f. That is, if the i-th entity in the entity vocabulary appears as a *positive* finding in the input EMR, then the i-th dimension of f is set to 1, and otherwise, it is set to 0. Moreover, the patient's age and gender are appended to f to get the hand-crafted feature for MLP. The MLP contains one dense layer activated by *sigmoid* function with 128 hidden units.

ECNN is trained with Adam optimizer (learning rate 0.001), 20 epochs and batch size of 32. The output of each convolutional tower and the output of the MLP are further concatenated before passing through the dropout and the softmax layer. Similar to Kim (2014), the dropout rate is empirically set to 0.5. A $|\mathbf{D}|$-dimensional feature is output by ECNN as the disease priors for the inference in the next where $\mathbf{D}$ is the disease set.

In ECNN, the CNNs are supposed to capture the *sequential signals* in the section texts and the MLP is supposed to encode the *feature of the critical entities*. By jointly modeling with CNNs and MLP, the proposed ECNN is expected to have superior performance than either of them alone.

3.3 Bayesian Network Ensembles

Although ECNN also outputs a probability distribution over all diseases, the result is not interpretable due to its end-to-end nature. However, the interpretability is very important in the CDS to explain how the diagnosis is generated by machines. Thus, we propose the Bayesian network ensembles on top of the output of ECNN to explicitly infer disease with PGMs. There are three steps:

3.3.1 Relation Extraction

We extract the relations between disease and other types of entities *(disease, finding)* where *finding* can be symptom, vital sign, test report finding, etc.

The rest of this paper will use *finding* to denote any type of entities other than *disease*. Relation extraction is performed in conjunction with the *(disease, finding)* co-occurrence mining and the deep extraction model (Shi et al., 2019) from the EMR documents and the textbooks [5]. Then, the pairs with high co-occurrences larger than a support (e.g. 5) are preserved. The extracted relations are reviewed by 10 certificated physicians. The invalid extracted relations which result from issues like incorrect recognition of entities or polarities by NER, the symptom caused by the secondary diagnosis but incorrectly paired with the first diagnosis, are removed before adding to the medical knowledge graph. Therefore, the relation *(disease, finding)* in the medical knowledge graph can, to some extent, be interpreted as: *disease* causes *finding*.

In our study, the pairs are mined from 275,797 EMR documents of two medical departments (Gynaecology and Respiration). On average, each disease of Gynaecology in our experiments is associated with 24 findings and that of Respiration is 42. For Gynaecology, there are 33 diseases, 305 symptoms, 143 vital signs and 25 test report findings in the PGMs. For Respiration, there are 21 diseases, 263 symptoms, 187 vital signs and 31 test report findings in the PGMs.

3.3.2 Relation Weights Estimation

We experiment with six classical text features as the relation weights in this study.

(1) Occurrence. The weight of finding i given disease j is:

$$w(i;j) = \frac{n(i,j)}{\sum_k n(k,j)}, \tag{1}$$

where $n(i,j)$ is the number of co-occurrences of finding i and disease j. $w(i;j)$ is computed by the type of findings.

(2) TF-IDF Feature. Similar to TF-IDF feature in information retrieval, the weight of finding i given disease j is:

$$w(i;j) = n(i,j) * (\log \frac{|\mathbf{D}|+1}{n_i+1} + 1), \tag{2}$$

where n_i is the number of diseases whose EMR documents contain finding i.

(3) TFC Feature. TFC feature (Salton and Buckley, 1988) is a variant of TF-IDF and it estimates the weight of finding i given disease j as:

$$w(i;j) = \frac{n(i,j) * \log \frac{|\mathbf{D}|}{n_i}}{\sqrt{\sum_k (n(k,j) * \log \frac{|\mathbf{D}|}{n_k})^2}}. \tag{3}$$

(4) TF-IWF Feature. The Term-Frequency Inverse-Word-Frequency (TF-IWF) feature (Basili et al., 1999) estimates the weight of finding i given disease j as:

$$w(i;j) = n(i,j) * (\log \frac{\sum_k t_k}{t_i})^2, \tag{4}$$

where t_i represents the number of occurrences of word i in the whole training corpus.

(5) CHI Feature. CHI feature (χ^2 Test) measures how much a term is associated with a class from a statistical view. The CHI feature of finding i given disease j is (Yang and Pedersen, 1997):

$$w(i;j) = \frac{N * (A * D - C * B)^2}{(A+C) * (B+D) * (A+B) * (C+D)}, \tag{5}$$

where N, A, B, C and D are the number of all documents, the number of documents containing finding i and belonging to disease j, the number of documents containing i but not belonging to j, the number of documents belonging to j but not containing i, and the number of documents not containing i and not belonging to j.

(6) Mutual Information. This feature assumes that the higher the strength between a finding and a disease, the higher their mutual information will be. Similar to the definition in CHI feature, this feature is defined as:

$$w(i;j) \approx \log \frac{A * N}{(A+C) * (A+B)}. \tag{6}$$

The above features are normalized by disease before applying to the diagnosis inference. By default, the average of the six features is used as the connection weight.

3.3.3 Diagnosis Inference

We propose the Bayesian network ensembles for the diagnosis inference. Specifically, a group of PGMs with the extracted relations and weights are ensembled towards the final predictions.

Firstly, multiple *bipartite graphs* between disease nodes and each type of finding nodes are derived from the medical knowledge graph. For M types of findings, there will be M bipartite graphs. In later experiments, $M = 3$, i.e. *(disease, symptom)*, *(disease, vital sign)* and *(disease, test result finding)*. Based on the findings extracted from EMR document, each bipartite graph can be independently used to infer the disease distribution.

[5] The undergraduate teaching materials in most of the medical schools in China, authorized by the publisher.

For Bayesian inference, we compute the posterior probability of diseases given the findings in the EMR document extracted by NER:

$$\Pr(d|F^+, F^-) = \frac{\Pr(d, F^+, F^-)}{\Pr(F^+, F^-)}, d \in \mathbf{D}, \quad (7)$$

where F^+ and F^- are the sets of the positive and the negative findings in the given EMR document, respectively. Following Eq. (7), it is straightforward to get $\Pr(d|F^+_{sym}, F^-_{sym})$, $\Pr(d|F^+_{sign}, F^-_{sign})$ and $\Pr(d|F^+_{test}, F^-_{test})$ w.r.t. the predictions based on symptom alone, vital sign alone and test report finding alone. To compute the joint probability $\Pr(d, F^+, F^-)$ and $\Pr(F^+, F^-)$, we refer the readers to the QuickScore method (Heckerman, 1990) and the deduction therein. To speed up computation when a disease is associated with too many positive findings, the variational method on the PGMs is applied (Jordan et al., 1999).

Next, we assemble these bipartite graphs in different ways to get three variants of PGMs (Fig. 1).

(1) **Parallel**. This method independently performs inference with each type of finding and average their results:

$$\begin{aligned}\Pr(d|F^+, F^-) = avg(&\Pr(d|F^+_{sym}, F^-_{sym}),\\ &\Pr(d|F^+_{sign}, F^-_{sign}), \Pr(d|F^+_{test}, F^-_{test})). \quad (8)\end{aligned}$$

Parallel assumes that the ways to diagnose disease are different using different types of entities, and their predictions can complement each other. An extension of **Parallel** is to perform a *weighted sum* of the three predictions. For simplicity concerns, we experiment with equal weights in this paper.

(2) **Universal**. This method mixes all types of findings together into a single network:

$$\begin{aligned}&\Pr(d|F^+, F^-) = \quad (9)\\ &\Pr(d|F^+_{sym}, F^-_{sym}, F^+_{sign}, F^-_{sign}, F^+_{test}, F^-_{test}).\end{aligned}$$

It means that **Universal** does not distinguish the types of entities and performs the type-free Bayesian inference. Compared with the other two PGM variants, the connections between diseases and findings in **Universal** are much denser. It assumes that the prediction benefits from the joint inference by *seeing* more findings of multiple types at the same time.

(3) **Cascade**. This method constructs the multi-layer Bayesian networks with finding types as layers and use the output of the previous layer as the prior probability for the current layer.

$$\begin{aligned}\Pr(d_{sym}) &= \Pr(d|F^+_{sym}, F^-_{sym})\\ &\quad s.t., d \sim \Pr(d_{CNN}),\\ \Pr(d_{sign}) &= \Pr(d|F^+_{sign}, F^-_{sign})\\ &\quad s.t., d \sim \Pr(d_{sym}),\\ \Pr(d_{BN}) &= \Pr(d_{test}) = \Pr(d|F^+_{test}, F^-_{test})\\ &\quad s.t., d \sim \Pr(d_{sign}), \quad (10)\end{aligned}$$

where $\Pr(d_{CNN})$ is the disease probability distribution computed by the convolutional networks in Sec. 3.2 and $d \sim \Pr(d_x)$ means that variable d satisfies *prior probability distribution* $\Pr(d_x)$. **Cascade** first infers disease with symptoms alone and uses the disease probability from ECNN as priors. Then, it infers disease with vital signs alone and uses the disease probability from symptom-based inference as priors. Finally, it infers disease with test report findings alone and uses the disease probability from the previous output as priors. We present the cascade appraoch in such order because it shows the best results compared to those in other orders in our experiments. **Cascade** assumes that each type of entities can be used to *refine* the previous predictions by incorporating additional information.

The output of the above three PGMs are ensembled, e.g. weighted sum, as the final predictions. In all, the proposed framework takes the raw EMR document and the NER results as input, and outputs the diagnosis predictions.

Although we experiment with three types of entities in this paper, the proposed Bayesian network ensemble method is not limited to these types of entities. It is easy to add more entity types in the proposed method when applicable.

3.4 The Interpretability of BN Ensembles

One of the major contributions of this work is to bring interpretability into automatic diagnosis by stacking the Bayesian network ensembles on top of the convolutional networks. We illustrate how the predictions are explained, i.e. *interpretability*, by BN with Fig. 2. We use the symptom-based bipartite graph to illustrate for the simplicity concern, and the other types of entities explain the predictions in the same way.

In Fig. 2, if only *pharyngalgia* is extracted from a patient's EMR, then *upper respiratory infection (URI)* will be predicted with high probability but the probability of *pneumonia* and *phthisis* will

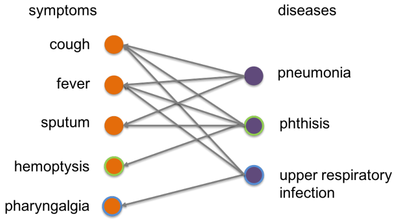


Figure 2: The example of the interpretability of Bayesian network. The connection from disease d to symptom s represents that d has some probability to cause s to be present. If d is diagnosed, the detected symptoms from EMR that are connected with d can be used to explain the diagnosis.

be set to the minimum because both of them are not likely to cause pharyngalgia based on their co-occurrences in the corpus. The proposed method can explain the prediction of URI with symptom pharyngalgia and their co-occurrence times besides the prediction probability.

If *pharyngalgia* and *hemoptysis* are both extracted from a patient's EMR, then URI as well as phthisis will be predicted with some positive probability (their rankings depend on both their prior probability and their connection weights to pharyngalgia and hemoptysis), but *pneumonia* will be predicted with the minimum probability. This is because the noisy-OR gate is used in the Bayesian inference (Heckerman, 1990). The proposed method explains the prediction of URI with the positive finding of symptom pharyngalgia and explains the prediction of phthisis with the positive finding of symptom hemoptysis as well as their co-occurrences.

4 Experiments and Results

In this section, we will introduce the data sets we experiment with and the evaluation results.

4.1 Data Sets

The proposed framework is evaluated on the real EMR documents (mostly admission records). We have collaborated with several top hospitals in China and we are authorized to conduct experiments with 275,797 EMR documents of two medical departments for the evaluation (see Table 3).[6]

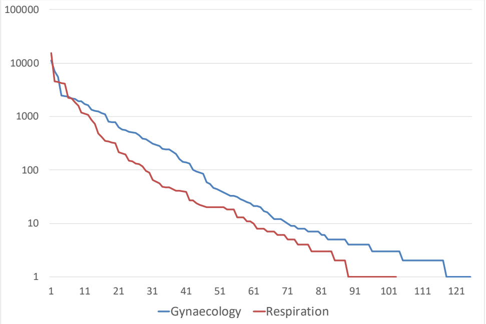


Figure 3: The long-tail distribution of diagnosis. The x-axis indexes the names of diagnosis. The y-axis counts the occurrences of diagnosis in the log scale.

Table 3: The statistics of the data sets. The table represents the document counts by source. # means *the number of*. "# collected" is the number of the collected EMR documents in the our experiments.

Departments	# collected	# test	# disease
Gynaecology	191,645	606	33
Respiration	84,152	214	21

The collected EMR documents are processed as follows: The main diagnosis in each EMR document is extracted as its disease label. Then, we select the top diseases from the collected EMR documents, which results in 33 diseases from Gynaecology (including Salpingitis, Cervical Carcinoma, Endometritis, Fibroid, etc) and 21 diseases from Respiration (including Upper Respiratory Infection, Chronic Bronchitis, Pneumonia, Asthma, Lung Cancer, etc) that cover over 90% of all EMR documents. There is a long-tail distribution of EMR documents by diseases as shown in Fig. 3, and each of the selected diseases has over 100 EMR documents for training. The other diseases are discarded in the experiments due to the lack of enough EMR documents to train a trustworthy model. Next, in order to ensure the validity of the disease labels in the test set, we recruit 10 professional physicians to review the labels by evenly sampling EMR documents under each disease. In this way, we collected 606 reviewed EMR documents for Gynaecology and 214 for Respiration as the test set (See disease distribution in supplemental files). The rest EMR documents are used for training. Since we are not given the identity of patient w.r.t. each EMR, the training and the testing sets are considered disjoint. In later experiments, we separately report the performance under both departments. It is more important and difficult to distinguish diseases within the same department than that across departments due to the overlapping symptoms, signs and test report findings among the similar diseases.

[6] Unfortunately, we have not yet obtained the permission from the hospitals to make the evaluation data sets public at this moment because EMR documents are legally protected by the Chinese laws and there is too much sensitive information about the patients and the doctors in them. We are currently working with the hospitals in contributing the benchmark EMR data sets for automatic diagnosis, but it takes time due to the legal issues. We suggest the readers to focus their attention on the contribution of the novel automatic diagnosis framework in this paper.

Table 4: The accuracy of the different diagnosis methods on two medical departments. Top-k sensitivity is used as the accuracy measurement.

Methods	Gynaecology		Respiration	
	Top-1	Top-3	Top-1	Top-3
CAML (2018)	58.6%	76.3%	60.7%	82.7%
CNN (2018)	61.0%	82.8%	61.7%	80.8%
ACNN (2018)	62.1%	83.3%	60.7%	84.6%
PGM-C	50.8%	64.6%	26.6%	47.6%
PGM-P	56.1%	69.3%	31.3%	45.3%
PGM-U	56.2%	69.6%	33.6%	57.9%
PGM-E	53.9%	70.2%	28.0%	48.1%
ECNN	68.9%	86.7%	65.8%	81.7%
ECNN-PGM-C	71.4%	88.6%	52.8%	82.7%
ECNN-PGM-U	72.9%	88.6%	59.3%	87.8%
ECNN-PGM-P	73.2%	88.4%	**68.2%**	87.3%
ECNN-PGM-E	**73.4%**	**88.8%**	64.0%	**88.3%**

4.2 Experimental Results

We conduct experiments on the collected data sets to evaluate the performance of the framework.

4.2.1 Experimental Settings

In the experiments, we used four CNN towers ($N = 4$) w.r.t. CC, HPI, PE and TR, and each tower has three channels with kernel length 3, 4 and 5 (representing 3-grams, 4-grams and 5-grams).

We use Jieba package[7] to perform Chinese word segmentation on the training set and remove the punctuation from the segmentation results. The segmented word corpus is used to train the 100-dimensional word embeddings using the Word2Vec (Mikolov et al., 2013) method (window as 5, min support as 5) implemented in the gensim package[8]. The top 100,000 frequent segmented words consist of the word vocabulary in the embedding layer of ECNN. Thus, the size of the embedding layer is (100000, 100).

Besides, the top 10,000 frequent entities (not segmented words) as well as age and gender are used to construct the one-hot feature into MLP which consists of one hidden dense layer (128 Sigmoid units) due to the efficiency consideration. Similar to Kim (2014), the dropout rate is empirically set to 0.5. By default, we use the average of all six relation weights in the experiments. The final predictions are the average of the three PGM variants. ECNN and PGMs are trained separately offline.

4.2.2 Performance Accuracy

Table 4 shows the Top-k sensitivity (The micro average of the per-disease Top-k sensitivity, commonly used as the accuracy measurement in healthcare studies (Liang et al., 2019).) under two departments. Generally, sensitivity is ususally used in binary classification (mostly output *yes* or *no*). Similarly, when we are dealing with classification of multi-class rather than binary classification, the proposed automatic diagnosis model outputs the probability distribution over K diseases (classes) for a given EMR. Suppose there are l_i out of n_i cases, where d_i is included in the Top-k predictions (ranked by probability) for the n_i EMRs of disease d_i. The Top-k sensitivity of the proposed model on disease d_i is: $\frac{l_i}{n_i}$. Furthermore, in the overall evaluation of the proposed model on all diseases, we use the micro average of all classes as the overall Top-k sensitivity:

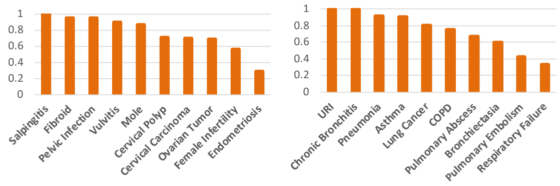
$$sensitivity = \frac{\sum_i l_i}{\sum_i n_i}. \tag{11}$$

CAML (Mullenbach et al., 2018) performs the label-wise attention on top of a CNN model. CNN (Yang et al., 2018) concatenates CC, HPI and TR together before sending to the multi-channel CNN model. ACNN (Girardi et al., 2018) incorporates the gram-level attention with a CNN model. The empirical settings of hyper parameters are selected from the original papers. Besides, they share the same training set, training epochs, learning rate and batch size with the proposed methods.

Among the proposed methods, *PGM-** (*-C*, *-P*, *-U* and *-E* represent *Cascade*, *Parallel*, *Universal* and *Ensemble*, respectively) are the methods that solely relies on the Bayesian networks which use the disease distribution in the training set as the prior probability. ECNN is the proposed method without the BN ensembles. *ECNN-PGM-** are the combined methods while *ECNN-PGM-E* is the proposed method with ECNN and Bayesian network ensembles in Figure 1. According to the results: (1) Most of the proposed methods *ECNN-PGM-** outperform the previous automatic diagnosis methods, which shows the effectiveness of the proposed methods. (2) ECNN outperforms CNN due to the incorporation of medical entities. Jointly modeling with free texts and medical entities brings extra accuracy performance compared with modeling with only either one. (3) Stacking Bayesian Networks on top of the neural networks is very likely to further improve the performance, especially with the ensemble of the predictions from multiple PGMs.

[7] https://github.com/fxsjy/jieba

[8] https://radimrehurek.com/gensim/



(a) Gynaecology (b) Respiration

Figure 4: Top-1 sensitivity by diseases.

4.2.3 Error Analysis

Fig. 4 shows the Top-1 sensitivity on some diseases. The performances across diseases are quite different. For example, the Top-1 sensitivity of Salpingitis is 100% but that of Endometriosis is 29% in the evaluation. Salpingitis can be identified by combining general symptoms and ultrasonic exam results. However, from the perspective of physicians, Endometriosis is difficult to diagnose by nature because it shares common symptoms like dysmenorrhea and irregular menstruation with other Gynecologic diseases. These shared findings misguide the classifier towards other similar diseases. Similarly, among the respiratory diseases, patients with Pulmonary Embolism, Respiratory Failure and Bronchiectasia share symptom *dyspnea* which makes it difficult to distinguish between them. In contrast, Upper Respiratory Infection (URI) is easy to diagnose because it causes throat pain and rhinorrhea unlike the other respiratory diseases.

Based on the analysis, the diagnosis performance of a disease is higher if it shares less findings with other diseases or it has more specific findings.

4.2.4 Interpretability

The interpretability is reflected on the observed findings in the EMR that connect to the predicted disease in the medical knowledge graph as well as their co-occurrences. We generate the prediction explanation with the following template: *The patient is diagnosed as disease d because (s)he is suffering from symptom s_i, and (s)he has the vital sign of v_j, and the lab test (or PACS report) shows (s)he has t_k. Besides, s_i, v_j and t_k have been found on the patients of d for n_i, n_j, n_k times, respectively, in the previous EMR documents that support this diagnosis.*

Since the extracted relations in the medical knowledge graph are reviewed by the certificated physicians, the validity of explanation is guaranteed from the clinical perspective. We randomly select 50 testing samples per department whose Top-1 diagnosis prediction is correct and generate the explanation for the diagnosis prediction with the above template. The explanation is evaluated by three certificated physicians. The evaluation is subjective, but all of them agree that the prediction is well-supported by the generated explanation.

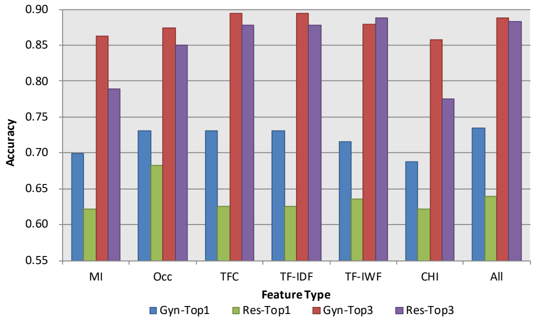


Figure 5: The accuracy of ECNN-PGM-E using different types of features. *Gyn* and *Res* represent *gynaecology* and *respiration*, respectively. MI and Occ are mutual information and occurrence, respectively.

4.2.5 Feature Importance

Figure 5 shows the accuracy performance using different types of features. We can see that in this evaluation, TFC, TF-IDF and the average of all features are likely to lead to higher accuracy compared to the other features where the accuracy of Top-3 prediction is over 88%.

In all, the above experiments prove that the proposed framework can improve the accuracy of automatic diagnosis and bring reasonable interpretability into the predictions in the same time.

5 Conclusion

In this paper, we investigate the problem of automatic diagnosis with EMR documents for clinical decision support. We propose a novel framework that stacks the Bayesian Network ensembles on top of the Entity-aware Convolutional Neural Networks. The proposed design brings interpretability into the predictions, which is very important for the AI-empowered healthcare, without compromising the accuracy of convolutional networks. The evaluation conducted on the real EMR documents from hospitals validates the effectiveness of the proposed framework compared to the baselines in automatic diagnosis with EMR.

Acknowledgement

We thank all the professional physicians led by Dr. Shi and Dr. Hu who have contributed in the annotation tasks in our experiments.

References

Padmanabhan Anandan, Yan Huang, Kazumi Nishikawa, BBorie Park, Eric S. Sullivan, Jingyu Wang, and Xu Shan. 2019. AI in health care: Capacity, capability, and a future of active health in Asia. *MIT Technology Review Insights*, pages 1–25.

Roberto Basili, Alessandro Moschitti, and Maria Teresa Pazienza. 1999. A text classifier based on linguistic processing. In *IJCAI Workshop on Machine Learning and Information Filtering*, Stockholm, Sweden.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2017. Multi-label classification of patient notes a case study on icd code assignment. In *AAAI Workshops*, pages 409–416.

Eta S. Berner. 2007. *Clinical Decision Support Systems*. Springer.

Jun Chen, Jingbo Zhou, Zhenhui Shi, Bin Fan, and Chengliang Luo. 2019. Knowledge abstraction matching for medical question answering. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 342–347.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016a. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016b. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*, pages 3504–3512.

Dai Dai, Xinyan Xiao, Yajuan Lyu, Shan Dou, Qiaoqiao She, and Haifeng Wang. 2019. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. In *AAAI*, Honolulu, Hawaii, USA.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25:24–29.

Ivan Girardi, Pengfei Ji, An phi Nguyen, Nora Hollenstein, Adam Ivankay, Lorenz Kuhn, Chiara Marchiori, and Ce Zhang. 2018. Patient risk assessment and warning symptom detection using deep attention-based neural networks. In *EMNLP Workshop*, pages 139–148, Brussels, Belgium.

David Heckerman. 1990. A tractable inference algorithm for diagnosing multiple diseases. *Machine Intelligence and Pattern Recognition*, 10:163–171.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In *ACL*, pages 1399–1408, Florence, Italy.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746—-1751, Doha, Qatar.

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, pages 364–379.

Christy Li, Dimitris Konomis, Graham Neubig, Pengtao Xie, Carol Cheng, and Eric Xing. 2017. Convolutional neural networks for medical diagnosis from admission notes. In *arXiv*.

Huiying Liang, Brian Y. Tsui, Hao Ni, Carolina C. S. Valentim, Sally L. Baxter, and et al. 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*, 25:433–438.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sanchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Qianlong Liu, Zhongyu Wei, Baolin Peng, Xiangying Dai, Huaixiao Tou, Ting Chen, Xuanjing Huang, and Kam fai Wong. 2018. Task-oriented dialogue system for automatic diagnosis. In *ACL*, pages 201—-207, Melbourne, Australia.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representation of words and phrases and their compositionality. In *NeurIPS*, pages 3111—-3119.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *NAACL*, pages 1101—1111, New Orleans, Louisiana, USA.

Aaditya Prakash, Siyuan Zhao, Sadid A. Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed memory networks for clinical diagnostic inferencing. In *AAAI*, pages 3274–3280.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Ying Sha and May D. Wang. 2017. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, pages 233–240, Boston, MA, USA.

Xue Shi, Yingping Yi, Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, Zongcheng Ji, Yaoyun Zhang, and Hua Xu. 2019. Extracting entities with attributes in clinical text via joint deep learning. *Journal of the American Medical Informatics Association*, pages 1584–1591.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, pages 412—-420, Nashville, TN, USA.

Zhongliang Yang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang, and Pengcheng Luo. 2018. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific Reports*, 8(6329).

Shiyue Zhang, Pengtao Xie, Dong Wang, and Eric P. Xing. 2017. Medical diagnosis from laboratory tests by combining generative and discriminative learning. In *arxiv*.