Identifying Risk Factors For Heart Disease in Electronic Medical Records: A Deep Learning Approach

Thanat Chokwijitkul[1], **Anthony Nguyen**[2], **Hamed Hassanzadeh**[2], **Siegfried Perez**[3]
[1]School of Information Technology and Electrical Engineering, The University of Queensland
[2]The Australian e-Health Research Centre, CSIRO
[3]Emergency Department, Logan Hospital
t.chokwijitkul@uqconnect.edu.au
{Anthony.Nguyen, Hamed.Hassanzadeh}@csiro.au
SiegfriedRobert.Perez@health.qld.gov.au

Abstract

Automatic identification of heart disease risk factors in clinical narratives can expedite disease progression modelling and support clinical decisions. Existing practical solutions for cardiovascular risk detection are mostly hybrid systems entailing the integration of knowledge-driven and data-driven methods, relying on dictionaries, rules and machine learning methods that require a substantial amount of human effort. This paper proposes a comparative analysis on the applicability of *deep learning*, a re-emerged data-driven technique, in the context of clinical text classification. Various deep learning architectures were devised and evaluated for extracting heart disease risk factors from clinical documents. The data provided for the 2014 i2b2/UTHealth shared task focusing on identifying risk factors for heart disease was used for system development and evaluation. Results have shown that a relatively simple deep learning model can achieve a high micro-averaged F-measure of 0.9081, which is comparable to the best systems from the shared task. This is highly encouraging given the simplicity of the deep learning approach compared to the heavily feature-engineered hybrid approaches that were required to achieve state-of-the-art performances.

1 Introduction

Heart disease is a leading cause of morbidity and mortality worldwide (Benjamin et al., 2017). As failure to recognise atypical representations of such serious illness may lead to adverse outcomes, accurate diagnosis is crucial to ensure that patients are placed on the proper treatment pathway. Electronic medical records (EMR) can be used to improve the diagnosis ability along with measuring the quality of care. The rapid adoption of EMRs along with the necessity to enhance the quality of health care has incentivised the development of natural language processing (NLP) in the medical domain. An abundant amount of clinical information used for medical investigation is organised in unstructured narrative form, which is suitable for expressing medical concepts or events but challenging for analysis and decision support as gaining a full aspect of a patients medical history by reading through EMRs is significantly time-consuming, especially when only a specific piece of information is needed. The difficulty of this process increases in the case of heart disease due to its complex progression, which regularly involves various factors including lifestyle and social factors as well as specific medical conditions (Stubbs and Uzuner, 2015). Various methods have been proposed in the field of clinical concept extraction, ranging from simple pattern matching to systems based on symbolic or statistical data and machine learning (Meystre et al., 2008; Gonzalez-Hernandez et al., 2017). Those previously proposed approaches have shown promising results but it is very difficult to reach that point due to the assiduous process of defining rules and extracting features. This is where deep learning comes in as this intriguing re-emerged concept can alleviate heavily human dependent efforts required for knowledge-based approaches and the lack of the ability of many conventional machine learning algorithms to learn without the necessity of careful feature engineering with considerable domain expertise (LeCun et al., 2015).

This paper presents a comparative analysis of two widely used deep learning architectures, namely convolutional neural network (CNN) and

Proceedings of the BioNLP 2018 workshop, pages 18–27
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

recurrent neural network (RNN) as well as three RNN variants, including long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), bidirectional long short-term memory (BLSTM), and gated recurrent unit (GRU) (Cho et al., 2014), for extracting cardiac risk factors from EMRs. Using the data set from the i2b2/UTHealth shared task (Stubbs and Uzuner, 2015), the goal is to determine the risk factor indicators contained within each document along with the temporal attributes with respect to the document creation time (DCT).

2 Related Work

2.1 Deep Learning for Clinical Information Extraction

Many recent publications have focused on extracting relevant clinical information from EMRs using deep learning. One of the most fundamental tasks involves the extraction of medical concepts from unstructured clinical notes. This concept extraction problem can be treated as a sequence labelling problem where the goal is to assign a clinically relevant tag to each word in an EMR (Jagannatha and Yu, 2016). Jagannatha and Yu (2016) experimented with different deep learning architectures based on recurrent networks, including GRUs, LSTMs and BLSTMs. It turned out that all the RNN variants outperformed the conditional random field (CRF) baselines, which had previously been considered the state-of-the-art method for information extraction in general.

As patient EMRs evolve over time, the sequentiality of clinical events can be used for disease progression analysis and the prediction of impending disease conditions (Cheng et al., 2016). Its temporality induces the necessity of assigning notions of time to each extracted medical concept. Fries (2016) devised a solution for such more complex problems by using a standard RNN initialised with word2vec (Mikolov et al., 2013a) vectors along with utilising DeepDive (Shin et al., 2015) for forming relationships and predictions. Li and Huang (2016) and Chikka (2016) also employed word embedding vectors within their frameworks but used CNNs to extract the temporal attributes instead. While still not state-of-the-art, these approaches produced competitive results in the field of temporal event extraction but also required a separate model for each subtask (extracting concepts and temporal attributes) and slight manual engineering (Shickel et al., 2017; Bethard et al., 2016). One thing to remark is that none of the existing systems has ever tried using a single, universal model that naturally learns the temporal characteristics of those concepts based on their contexts and incorporates them into the feature learning process, which can be used for extracting medical concepts and temporal attributes simultaneously. This work intends to explore this idea and prove that the aforementioned capability is well within the reach of deep learning.

2.2 i2b2/UTHealth Shared Task

In 2014, the Informatics for Integrating Biology and the Bedside (i2b2) issued an NLP shared task focusing on identifying risk factors for heart disease in clinical narratives. According to Stubbs et al. (2015), a total of 49 systems from 20 teams were submitted. The systems varied broadly, from rule-based systems to complex hybrid systems with a combination of machine learning techniques. Nevertheless, some similarities were found among the top systems including the use of preprocessing tools to obtain syntactic information and section headers for determining temporal labels. The results revealed that the top 10 systems achieved micro-averaged F1 scores over 0.87 while the top 6 systems were able to reach micro-averaged F1 scores over 0.90. The most successful system managed to achieve an F1 score of 0.928 (Roberts et al., 2015) while the averaged F1 score among all the systems was 0.815. While half of the top 10 teams used a combination of knowledge-driven methods, such as lexicon and rules, and machine learning algorithms, including CRF, support vector machine (SVM), Naïve Bayes classifier and Maximum Entropy, none of the participants attempted to integrate neural networks or deep learning into their systems. Furthermore, there has not existed any approaches that use deep learning to extract risk factor indicators from the shared task data since its inception in 2014, which is a research gap that this work intends to fill.

3 Methodology

3.1 Dataset

The dataset used in this work is the corpus provided for the 2014 i2b2/UTHealth shared task. The corpus consists of 1,304 medical records describing 296 diabetic patients for cardiovascular risk factors and time attributes with respect to the DCT. The dataset was split by the challenge

Risk Factor	Indicator	Training Instances	Testing Instances	Time Attribute
CAD	mention, event, test, symptom	1186	784	✓
Diabetes	mention, high A1c, high glucose	1695	1180	✓
Obesity	mention, high BMI	433	262	✓
Hyperlipidemia	mention, high cholesterol, high LDL	1062	751	✓
Hypertension	mention, high blood pressure	1926	1293	✓
Medication	ACE inhibitor, amylin, anti-diabetes, ARB, aspirin, beta blocker, calcium channel blocker, diuretic, DPP4 inhibitors, ezetimibe, fibrate, GLP1 agonist, insulin, Meglitinide, metformin, niacin, nitrate, obesity medications, statin, sulfonylurea, thiazolidinedione, thienopyridine	8638	5674	✓
Smoking	current, past, ever, never, unknown	771	512	n/a
Family history	present, not present	790	514	n/a

Table 1: The indicators associated with each cardiac risk factor and the number of training and testing instances at annotation level

Evidence Type	Example
Phrase-based	Significant PMH for **CAD**, **HTN**, GERD, and past cerebral embolism
Logic-based	Seen in Cardiac rehab locally last week and **BP 170/80**
Discourse-based	**Findings suggestive of an obstructive, coronary lesion in the left circumflex distribution**

Table 2: Three types of evidence

Risk Factor	Phrase-based	Logic-based	Discourse-based
CAD	mention	n/a	event, test, symptom
Diabetes	mention	high A1c, high glucose	n/a
Obesity	mention	BMI	n/a
Hyperlipidemia	mention	high cholesterol, high LDL	n/a
Hypertention	mention	high blood pressure	n/a
Medication	all types	n/a	n/a
Smoking	n/a	n/a	all statuses
Family history	n/a	n/a	all statuses
Percentage of training instances	85.33%	8.10%	6.57%

Table 3: Relationships between the indicators and evidence types and the percentage of training instances belonging to each type

provider. The training set consists of 60% of the entire dataset (790 records) and the test set contains the remaining 40% (514 records). The annotation guidelines describe a set of annotations to indicate the presence of diseases (*coronary artery disease (CAD)* and *diabetes*), relevant risk factors (*hyperlipidaemia, hypertension, obesity, smoking status* and *family history*) and associated medications. Each annotation for a risk factor also has an indicator value from its own set (see Table 1) as well as the time attribute (*before, during* or *after* the DCT). Figure 1 shows an example of annotations used for training and evaluation. The ultimate goal is to classify risk factors and time indicators at document level as per Gold Standard annotation.

The evidence of risk factor indicators can be categorised into three types according to the terminologies described by Chen et al. (2015), which include phrase-based, logic-based and discourse-based indicators as presented in Table 2. Phrase-based indicators are those that can be identified directly by locating relevant phrases or particular names. Logic-based indicators are indirect information that needs a comparison or further analysis after being identified. Finally, discourse-based indicators are those that appear in the form of sentences and may require a parsing process. The relationships between indicators and evidence types are listed in Table 3.

3.2 Problem Formation and Evaluation

The classification of risk factors and time indicators was posed as a document-level classification problem. This can be seen as a multilabel classification task where multiple labels are identified given an EMR. However, unique to the annotation guideline (Stubbs and Uzuner, 2015) and

Complete version (for training):
<DIABETES start="122" end="130" text="diabetes" time="before DCT" indicator="mention"/>
<DIABETES start="512" end="528" text="diabetes type II" time="before DCT" indicator="mention"/>
<DIABETES start="701" end="718" text="diabetes mellitus" time="before DCT" indicator="mention"/>

Gold standard version (for evaluation):
<DIABETES time="before DCT" indicator="mention"/>

Figure 1: Each complete annotation contains token-level information (risk factor tag, risk factor indicator, offset, text information, and time attribute) while each gold standard annotation contains document-level information (risk factor tag, risk factor indicator and time attribute) and cannot be duplicated.

the structure of the training data, which contains phrase-level risk factor and time indicator annotations (see Figure 1), *it seems appropriate to formulate the problem as an information extraction task instead*. This approach regards data as a sequence of tokens labelled using the Inside-Outside (IO) scheme: *I* represents named entity tokens and *O* indicates non-entity ones. As the main goal is to determine the risk factor indicators contained within the record along with the temporal categories of those indicators with respect to the DCT, each entity is tagged with a label using the following format:

I-risk_factor.indicator.time

Figure 2 shows a sample EMR (represented by a sequence of words) and associated labels. In this example, the word "coronary" with the label "I-cad.mention.before_dct" can be interpreted that as a mention of CAD which was present before the document creation time.

Words:	he, has, coronary, artery, disease, and, diabetes
Labels:	O, O, I-cad.mention.before_dct, I-cad.mention.before_dct, I-cad.mention.before_dct, O, I-diabetes.mention.before_dct

Figure 2: A sample phrase in an EMR and associated labels

Given an EMR as input, the output is a sequence of labels, with each label belonging to a given word. After removing duplicate labels, *the final output will be a set of unique labels identified for that record* (excluding the O label). For the example in Figure 2, the final output will be generated as a set of two unique labels, including "I-cad.mention.before_dct" and "I-diabetes.mention.before_dct". These labels will be used to generate system annotations similar to the one presented in Figure 1 which will subsequently be evaluated against the gold standard annotations provided by the challenge provider using the micro-averaged recall, precision and F-measure as the primary evaluation metrics[1].

3.3 Deep Neural Network Models

3.3.1 Convolutional Neural Network

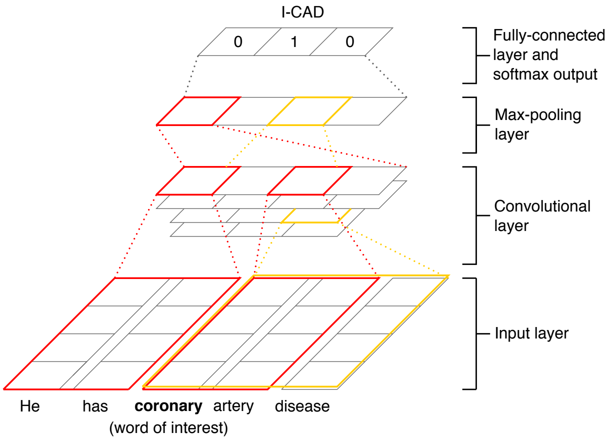

Figure 3: CNN architecture with multiple filter region sizes

The CNN model, as shown in Figure 3, is based on the CNN architecture of Kim (2014) but uses the window approach for NER, introduced by Collobert et al. (2011), to classify each individual word at a time instead of the entire sentence. This approach assumes the label of a word is dependent on its neighbouring words. Given a word to tag, a fixed size window of n words around the target word where n is odd is taken into account. A window of n words is represented as a matrix $\mathbf{S} \in \mathbb{R}^{d \times n}$:

$$\mathbf{S} = \begin{bmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_{n-\frac{(n-1)}{2}} & \cdots & \mathbf{w}_n \end{bmatrix} \quad (1)$$

[1] The official evaluation script is available at https://github.com/kotfic/i2b2_evaluation_scripts

where $\mathbf{w}_i \in \mathbb{R}^d$ is the d-dimensional word vector representing the ith word in $\mathbf{S}$ and $\mathbf{w}_{n-\frac{(n-1)}{2}}$ is the target word. Let $\mathbf{w}_{i:i+j}$ be the concatenation of words $\mathbf{w}_i, \mathbf{w}_{i+1}, ..., \mathbf{w}_{i+j}$. A convolution operation involves applying a *filter* $\mathbf{k} \in \mathbb{R}^{d \times h}$ to a window of h words, where $h < n$, to generate a new feature. For instance, a feature x_i is computed by

$$x_i = f(\mathbf{k} \cdot \mathbf{w}_{i:i+h-1} + b) \qquad (2)$$

where f is an activation function and $b \in \mathbb{R}$ is a bias. Note that this CNN architecture can employ multiple filter region sizes for extracting multiple features. This operation is applied to every possible window of words in the sequence $\{\mathbf{w}_{1:h}, \mathbf{w}_{2:h+1}, ..., \mathbf{w}_{n-h+1:n}\}$ to generate a *feature map* $\mathbf{x} = (x_1, x_2, ..., x_{n-h+1})$ where $\mathbf{x} \in \mathbb{R}^{n-h+1}$. The pooling layer then applies the max-pooling operation to down-sample each feature map by taking the maximum value $\hat{x} = \max(\mathbf{x})$ which represents the most important feature. Finally, multiple down-sampled feature maps form a fully-connected layer, which is used as inputs to the softmax distribution over all classes. The subsampled feature maps provide a sequence representation for softmax to map to an appropriate class.

3.3.2 Recurrent Neural Network

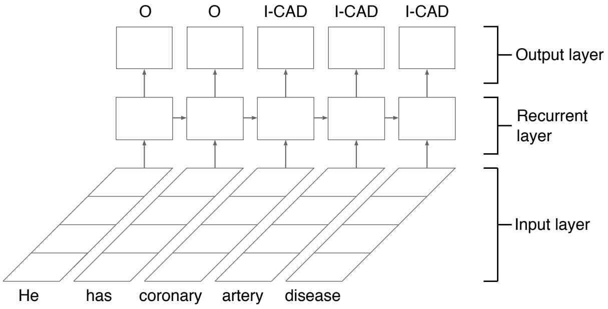

Figure 4: Basic structure of an RNN

A recurrent neural network is a class of neural networks specialised for processing sequential data. Unlike the CNN, the RNN uses a recurrent layer to learn the representation of clinical text, as shown in Figure 4. The input to an RNN is a word sequence of length l representing the *entire document*, denoted by a matrix $\mathbf{S} \in \mathbb{R}^{d \times l}$:

$$\mathbf{S} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & ... & \mathbf{w}_l \end{bmatrix} \qquad (3)$$

where $\mathbf{w}_i \in \mathbb{R}^d$ is the d-dimensional word vector representing the ith word in $\mathbf{S}$. In an *Elman-type network* (Elman, 1990), a hidden state output $\mathbf{h}_i$ is a result of nonlinear transformation of an input vector $\mathbf{w}_i$ and the previous hidden state $\mathbf{h}_{i-1}$:

$$\mathbf{h}_i = f(\mathbf{h}_{i-1}, \mathbf{w}_i) \qquad (4)$$

where f is a recurrent unit, such as a standard recurrent unit, LSTM and GRU. Finally, the hidden state $\mathbf{h}_i$ is then used as an input to softmax for identifying a risk factor in the IO format.

Bidirectionality. A bidirectional recurrent neural network (Schuster and Paliwal, 1997) consists of two separated recurrent layers for computing the forward hidden states $(\overrightarrow{\mathbf{h}_1}, \overrightarrow{\mathbf{h}_2}, ..., \overrightarrow{\mathbf{h}_l})$ and the backward hidden states $(\overleftarrow{\mathbf{h}_1}, \overleftarrow{\mathbf{h}_2}, ..., \overleftarrow{\mathbf{h}_l})$. In this settings, $\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_i}$ can be regarded as preserved information from the past and the future respectively. By using the hidden states from both directions combined, the network has complete past and future context for every point in the input sequence.

3.4 Pre-trained Word Embeddings

Due to the incapability of neural networks to process text input, each word is fed to the network as an index taken from a finite dictionary. As this simple representation does not contain much semantic information, the first layer of each network maps each index into its vector representation using pre-trained word embeddings. The pre-trained vectors were trained on the 2014 i2b2 dataset. The number of embedding dimensions was determined empirically. Given a small vocabulary (36,663 words) and a range of embedding dimensions from 20 to 300, an embedding dimension of 20 yielded best results. Each vector was trained via the word2vec's continuous bag-of-words (CBOW) model (Mikolov et al., 2013b) similar to that used by Kim (2014).

3.5 Hyperparameters and Training

The CNN model used 5-gram of each EMR as input since a window of 5 words has shown to be effective for many NLP tasks (Collobert et al., 2011). Based on the hyperparameters described by Kim (2014) and Zhang and Wallace (2015), the convolutional layer uses multiple filter region sizes $\{2, 3, 4\}$, each of which has 32 filters, and a rectifier (ReLU) as the activation function. For the RNN approach, experiments were performed on the standard RNN as well as its variants: LSTM, BLSTM and GRU. All the recurrent networks use the hyperbolic tangent as activation functions as it was considered one of the most common choices for RNN-type networks (Graves, 2012).

The hyperparameters apart from the above mentioned were tuned on the validation set (20% of the training set) using the hyperparameter tuning library within the framework of Bayesian optimisation, namely *Hyperopt* (Bergstra et al., 2013). Based on the hyperparameter optimisation results, all the networks were trained with mini-batch stochastic gradient descent using *Nadam* (Adam RMSprop with Nesterov momentum) (Dozat, 2016) with a batch size of 32. Dropout regularisation was also applied to the penultimate layer of each network for overfitting prevention. The resulting optimal values of other hyperparameters, including the number of hidden units (hidden), learning rate (lr), dropout rate and the number of epochs are listed in Table 4.

	CNN	RNN	GRU	LSTM	BLSTM
hidden	256*	256	256	512	256
lr	0.001	0.002	0.002	0.002	0.004
dropout	0.2	0.3	0.1	0.3	0.5
epochs	15	40	50	45	40

* The number of units in the fully connected layer

Table 4: Hyperparameters estimated by Hyperopt

4 Results and Discussion

4.1 Overall Performance

Results for each deep learning model's best run against state-of-the-art models from the 2014 i2b2/UTHealth shared task are listed in Table 5. Among the deep learning approaches, RNN-type networks outperformed CNN in the context of clinical text classification. Although the CNN model achieved the highest recall, its precision is far from being competitive, which results in a relatively low F-measure. A comparison between the RNN-type models shows that BLSTM achieved the highest micro-averaged F-measure (0.9081) on the test data, followed closely by GRU and LSTM. A two-tailed unpaired t-test was also performed to determine the significance of the difference in F-measure between the two best-performing networks. Over 50 independent training and testing sessions with different weight initialisation (drawn from the uniform distribution), the test yielded a statistically significant difference between the performance of BLSTM ($\mu = 0.903$, $\sigma = 0.002$) and GRU ($\mu = 0.899$, $\sigma = 0.002$) with $p < 0.05$, which implies that the improvement in performance of the BLSTM model is also statistically significant compared with that of other remaining models.

In comparison with the top performing systems from the previous work, the results reveal that the BLSTM model without employing any knowledge-driven approaches ranked in the top 6 systems, and was substantially better than the overall average (0.815) of all the participating systems in the shared task. As a universal classifier, the performance of the BLSTM model is auspicious since it produced only 0.0195 loss in F-measure when comparing against the first-ranked system (Roberts et al., 2015) which involves the use of a series of SVMs along with a rule-based classifier and additional annotations. Besides the best-performing model, the LSTM and GRU models ranked in the top 7 systems while the CNN and standard RNN models performed well within the top 10 systems from the shared task. This outcome concludes that simple deep learning models still can rank within the top 10 heavily feature-engineered best-performing systems from the shared task.

Model	Recall	Precision	F-score
BLSTM	0.9180	0.8983	**0.9081**
GRU	0.9091	**0.9002**	0.9046
LSTM	0.9191	0.8836	0.9010
RNN	0.8956	0.8844	0.8900
CNN	**0.9245**	0.8383	0.8793
Roberts et al. (2015)*	**0.9625**	0.8951	**0.9276**
Chen et al. (2015)*	0.9436	**0.9106**	0.9268
Torii et al. (2014)*	0.9409	0.8972	0.9185
Cormack et al. (2015)†	0.9375	0.8975	0.9171
Yang and Garibaldi (2014)*	0.9488	0.8847	0.9156
Shivade et al. (2015)†	0.9261	0.8907	0.9081
Chang et al. (2015)*	0.9387	0.8594	0.8973
NCU‡	0.9256	0.8586	0.8909
Karystianis et al. (2015)†	0.9007	0.8557	0.8776
Khalifa and Meystre (2015)†	0.8951	0.8552	0.8747

* A combination of knowledge- and data-driven approaches (hybrid)
† Knowledge-driven approaches only e.g. lexicon and rules
‡ Unknown (National Central University did not submit a paper)

Table 5: Experimental results and state-of-the-art systems from 2014 i2b2/UTHealth shared task

4.2 Performance on Individual Risk Factors

Table 6 shows the performance of the deep learning models on individual risk factors. All five architectures achieved micro-averaged F-measures

	CNN	RNN	GRU	LSTM	BLSTM
CAD	0.6553	0.7966	0.7972	0.8010	**0.8074**
Diabetes	0.9133	0.9227	0.9177	**0.9272**	0.9171
Obesity	0.8717	0.8739	0.8819	**0.8880**	**0.8880**
Hyperlipidemia	0.9154	0.9209	0.9100	0.9243	**0.9323**
Hypertension	0.8839	0.9093	0.9102	0.9043	**0.9187**
Medication	0.9075	0.8901	**0.9192**	0.9090	0.9171
Smoking	0.8350	0.8077	0.8146	0.8152	**0.8409**
Family history	0.9397	**0.9630**	0.9572	0.9591	**0.9630**
Overall	0.8798	0.8900	0.9046	0.9010	**0.9081**

Table 6: Micro-averaged F-measure for individual risk factor categories (best runs); highest F-measures for each category are bolded

over 0.87. These deep networks performed best on the family history category, achieved F-measures above 0.90 for the hyperlipidemia and diabetes risk factors, and maintained F-measures over 0.87 for the hypertension and obesity risk factors along with relevant medications. The worst classification performance of all the models was obtained for the CAD risk factor, followed by the smoking status.

Among the deep learning models, highest micro-averaged F-measures for most of the risk factor categories were achieved by the BLSTM network while the top performance for the diabetes and medication categories were obtained by the LSTM and GRU networks respectively. Lowest classification scores for most of the risk factor categories were achieved by the CNN model, which implies its inferiority in comparison with the RNN-type models for extracting cardiac risk factor information from EMRs. The overall outcome also reveals that even though the neural network architectures with the integration of recurrent units can be potentially applied to this particular task with higher success rate, the capability of the standard RNN is far from being highly efficient and thus using the gating mechanism as well and introducing bidirectionality can substantially increase the chance of achieving better performances.

4.3 Performance on Individual Risk Factor Indicators

The results in Table 7 reveals that phrase-based indicators have comparatively high F-measures in all models. As the deep learning approach for clinical concept extraction can be posed as a standard the named entity recognition task, specific keywords play a significant role in identifying named entities and an increase in the predictive performance is simply due to a tremendous amount of sample instances in the training data.

In contrast, the logic-based and discourse-based indicators have substantially lower F-measure. As both types of indicators infrequently appear in the training data (see Table 3), the primary cause of poor performance is likely due to the sparsity and imbalance of training instances.

	CNN	RNN	GRU	LSTM	BLSTM
Phrase-based	0.7679	0.6818	0.7810	0.7342	0.7808
Logic-based	0.3643	0.1857	0.2185	0.2114	0.2640
Discourse-based	0.5341	0.4983	0.5425	0.5328	0.5721

Table 7: The average of F-measure performances across all risk factor indicators for each evidence type

4.4 Error Analysis

4.4.1 Complex Textual Evidence

Even though phrase-based evidence may vary (e.g. CAD can appear as “heart disease” or “CAD”), these phrases along with a sufficiently large amount of samples are generally enough for deep neural networks to achieve high classification accuracy. However, the context of discourse-based evidence may appear to be as complex as “probable inferior and old anteroseptal myocardial infarction” or “Cath (5/88): 3v disease: RCA 90%, LAD 30% mid, 80% distal, D1 70%, D2 40% and 60%, LCx 30%, OM2 80%”. The difficulty of learning the patterns and identifying these indicators implies the need for a higher amount of training instances and perhaps amended semantic matching of medical terms to medical terminology resources such as the UMLS Metathesaurus

(Bodenreider, 2004) or Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT) (Stearns et al., 2001), such that information in EMRs can be more accurately extracted using deep learning.

4.4.2 Conditional Textual Evidence

Although deep learning requires less human effort and time than dictionary-based and rule-based approaches as it can automatically learn the patterns in data which results in more flexible predictive power, the experimental results demonstrate the limitation of such data-driven approach as it is infeasible to accurately identify logic-based indicators in the test set without having seen the numbers and their contexts in the training set. For example, it is unlikely for deep learning models to classify the evidence "glucose 420" as the diabetes.glucose indicator without learning that particular pattern during training as it is unable to perform comparison during classification whether 420 is greater than 126 (the glucose level greater than 126 is considered a risk factor (Stubbs and Uzuner, 2015)). A decrease in classification accuracy is primarily due to a massive amount of unforeseen evidence in the test data i.e. many numbers that imply the risk of heart disease never appear in the training set. In this case, utilising dictionaries and rules based on the domain knowledge would be more optimal than collecting more data in which every possible pattern, which may include every number that is considered a risk factor as well as its context, is required.

4.4.3 Data Sparsity and Class Imbalance

Figure 5 illustrates the relationship between classification performance of the BLSTM network[2] and the number of training instances in terms of risk factor indicators. When the number of samples is low (less than approximately 200 samples), each network's performance significantly varies depending on risk factor indicator. However, the prediction capability raises and tends to be more stable as the number of training instance increases. As many of the machine learning algorithms greatly suffer from insufficient and imbalanced data where the classes are not equally presented, it is not surprising if deep learning is severely impacted by the same problem. Inadequate training samples typically result in failure of pattern recognition while imbalanced classes in the training set tend to bias the trained models towards more common classes. These non-trivial issues likely explain the relatively poor classification results for various risk factor indicators, especially those that belong to the logic-based and discourse-based types, due to misclassification of either indicators or time attributes or both. Regarding the report from the 2014 i2b2/UTHealth risk factor challenge (Stubbs et al., 2015), all the participating systems also produced similar sets of results due to these problems.

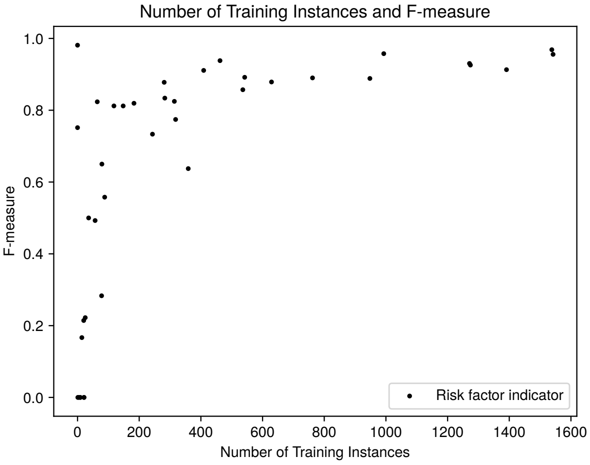

Figure 5: Effect of training-sample size illustrated by the relationship between classification performance of the BLSTM network and the number of training instances (risk factor indicator-level)

5 Conclusion

This work empirically evaluated the performance of different deep learning architectures for identifying risk factors for heart disease in clinical text. The experimental results showed that the deep learning approaches were not only comparable to highly feature-engineered hybrid systems but most importantly achieved relatively high performances without the help of any knowledge-driven methods. The findings leads to an anticipation that leveraging knowledge-based approaches with the BLSTM model could potentially provide significant performance improvements over best systems for extracting key cardiac risk factors from EMRs.

[2] The relationship between classification performance of the BLSTM network and the number of training instances is selected as it is the best-performing model from the experiment and the patterns found among other deep learning architectures are very similar.

References

Emelia J Benjamin, Michael J Blaha, Stephanie E Chiuve, Mary Cushman, Sandeep R Das, Rajat Deo, Sarah D de Ferranti, James Floyd, Myriam Fornage, Cathleen Gillespie, et al. 2017. Heart disease and stroke statistics 2017 update: a report from the american heart association. *Circulation*, 135(10):e146–e603.

James Bergstra, Dan Yamins, and David D Cox. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*, pages 13–20.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(1):D267–D270.

Nai-Wen Chang, Hong-Jie Dai, Jitendra Jonnagaddala, Chih-Wei Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2015. A context-aware approach for progression tracking of medical concepts in electronic medical records. *Journal of Biomedical Informatics*, 58:S150–S157.

Qingcai Chen, Haodi Li, Buzhou Tang, Xiaolong Wang, Xin Liu, Zengjian Liu, Shu Liu, Weida Wang, Qiwen Deng, Suisong Zhu, et al. 2015. An automatic system to identify heart disease risk factors in clinical texts over time. *Journal of Biomedical Informatics*, 58:S158–S163.

Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 432–440. SIAM.

Veera Raghavendra Chikka. 2016. CDE-IIITH at SemEval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1237–1240.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

James Cormack, Chinmoy Nath, David Milward, Kalpana Raja, and Siddhartha R Jonnalagadda. 2015. Agile text mining for the 2014 i2b2/UTHealth cardiac risk factors challenge. *Journal of Biomedical Informatics*, 58:S120–S127.

Timothy Dozat. 2016. Incorporating nesterov momentum into adam. *4th International Conference on Learning Representations (ICLR 2016)*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Jason Alan Fries. 2016. Brundlefly at SemEval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction. *arXiv preprint arXiv:1606.01433*.

G Gonzalez-Hernandez, A Sarker, K O'Connor, and G Savova. 2017. Capturing the patients perspective: a review of advances in natural language processing of health-related text. *Yearbook of Medical Informatics*, 26(01):214–227.

Alex Graves. 2012. *Supervised sequence labelling with recurrent neural networks*. Springer, Berlin, Heidelberg.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Abhyuday N Jagannatha and Hong Yu. 2016. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 856.

George Karystianis, Azad Dehghan, Aleksandar Kovacevic, John A Keane, and Goran Nenadic. 2015. Using local lexicalized rules to identify heart disease risk factors in clinical notes. *Journal of Biomedical Informatics*, 58:S183–S188.

Abdulrahman Khalifa and Stéphane Meystre. 2015. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics*, 58:S128–S132.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Peng Li and Heng Huang. 2016. University of Texas at Arlington (UTA) with deep learning based natural language processing (DLNLP) at SemEval-2016 task 12: deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. In

Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 1268–1273.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook of Medical Informatics 2008*, 35(128):128–144.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Kirk Roberts, Sonya E Shooshan, Laritza Rodriguez, Swapna Abhyankar, Halil Kilicoglu, and Dina Demner-Fushman. 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *Journal of Biomedical Informatics*, 58:S111–S119.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*.

Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. 2015. Incremental knowledge base construction using DeepDive. *Proceedings of the VLDB Endowment*, 8(11):1310–1321.

Chaitanya Shivade, Pranav Malewadkar, Eric Fosler-Lussier, and Albert M Lai. 2015. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *Journal of Biomedical Informatics*, 58:S103–S110.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2. *Journal of Biomedical Informatics*, 58:S67–S77.

Amber Stubbs and Özlem Uzuner. 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 58:S78–S91.

Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T Wiley, Daniel Zisook, and Yang Huang. 2014. De-identification and risk factor detection in medical records. In *Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data*.

Hui Yang and Jonathan Garibaldi. 2014. Automatic extraction of risk factors for heart disease in clinical texts. *Proceeding of the i2b2/UTHealth NLP Challenge*.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.