**Differentiation of lumbar disc herniation and lumbar spinal stenosis using natural language processing–based machine learning based on positive symptoms**

*GuanRui Ren, MD,1 Kun Yu, MS,2 ZhiYang Xie, PhD,1 Lei Liu, PhD,1 PeiYang Wang, MD,1 Wei Zhang, MS,1 YunTao Wang, PhD,1 and XiaoTao Wu, MD1*

1Zhongda Hospital, Medical College, Southeast University, Nanjing; and 2Nanjing Jiangbei Hospital, Nanjing, Jiangsu, China

**OBJECTIVE** The purpose of this study was to develop natural language processing (NLP)–based machine learning algorithms to automatically differentiate lumbar disc herniation (LDH) and lumbar spinal stenosis (LSS) based on positive symptoms in free-text admission notes. The secondary purpose was to compare the performance of the deep learning algorithm with the ensemble model on the current task.

**METHODS** In total, 1921 patients whose principal diagnosis was LDH or LSS between June 2013 and June 2020 at Zhongda Hospital, affiliated with Southeast University, were retrospectively analyzed. The data set was randomly divided into a training set and testing set at a 7:3 ratio. Long Short-Term Memory (LSTM) and extreme gradient boosting (XGBoost) models were developed in this study. NLP algorithms were assessed on the testing set by the following metrics: receiver operating characteristic (ROC) curve, area under the curve (AUC), accuracy score, recall score, F1 score, and precision score.

**RESULTS** In the testing set, the LSTM model achieved an AUC of 0.8487, accuracy score of 0.7818, recall score of 0.9045, F1 score of 0.8108, and precision score of 0.7347. In comparison, the XGBoost model achieved an AUC of 0.7565, accuracy score of 0.6961, recall score of 0.7387, F1 score of 0.7153, and precision score of 0.6934.

**CONCLUSIONS** NLP-based machine learning algorithms were a promising auxiliary to the electronic health record in spine disease diagnosis. LSTM, the deep learning model, showed better capacity compared with the widely used ensemble model, XGBoost, in differentiation of LDH and LSS using positive symptoms. This study presents a proof of concept for the application of NLP in prediagnosis of spine disease.

https://thejns.org/doi/abs/10.3171/2022.1.FOCUS21561

**KEYWORDS** natural language processing; machine learning; Long Short-Term Memory network; lumbar disc herniation; lumbar spinal stenosis; electronic health record

Approximately 266 million patients (3.63%) with lumbar degenerative disease are diagnosed yearly, which is a significant cause of disability worldwide.1 Lumbar disc herniation (LDH) and lumbar spinal stenosis (LSS) are two main types of lumbar degenerative disease. Neurogenic claudication, the common symptom of degenerative LSS, is the pain or discomfort that radiates into the buttocks, thigh, and lower leg.2 Lumbosacral radicular syndrome is typically characterized by radiating pain in the dermatome of a lumbar or sacral spinal nerve root and is most commonly caused by LDH.4 Despite differences in presentation and clinical course, distinguishing between these two causes of radiating lower-extremity pain is sometimes difficult, due to great variation in symptom patterns among patients.5 Rainville and Lopez found that patients with LSS had significantly more medical comorbidities, less-intense leg pain, and lower Oswestry Disability Index scores than those with LDH.5 Leg pain associated with LDH was more frequent in the anterior thigh and shin, and that of LSS in the posterior knee.5 It is essential to develop models to distinguish LSS and LDH as auxiliary tools for elevating diagnostic accuracy.

As a branch of artificial intelligence (AI), natural language processing (NLP) is an automated extraction of structured data from unstructured free-text data. The technology has been used in various industries to efficiently

**ABBREVIATIONS** AI = artificial intelligence; AUC = area under the curve; EHR = electronic health record; LDH = lumbar disc herniation; LSS = lumbar spinal stenosis; LSTM = Long Short-Term Memory; NLP = natural language processing; RNN = recurrent neural network; ROC = receiver operating characteristic; XGBoost = extreme gradient boosting.


**INCLUDE WHEN CITING** DOI: 10.3171/2022.1.FOCUS21561.

* G.R. and K.Y. contributed equally to this work and share first authorship.
abstract data from large repositories of text. With clinical notes remaining the standard of communication and documentation, the medical field is ideally suited to leverage informatics technology. In previous research, NLP with machine learning could discriminate excellently between cases of severe and nonsevere chest injury using the first 8 hours of clinical documentation. We hypothesized that NLP-based machine learning algorithms could “learn” from the big data of electronic health records (EHRs) and, thus, provide as accurate a diagnosis as possible, offering a promising auxiliary tool in the spine domain. However, to the best of our knowledge, there are no NLP algorithms for diagnosing spine diseases from EHRs. Our purpose was to develop NLP algorithms to automatically differentiate LDH and LSS based on positive symptoms in free-text admission notes.

Recurrent neural networks (RNNs) are a kind of deep learning algorithm widely adopted for sequential data such as text, audio, and video. RNN layers consist of recurrent cells whose states are affected by both past states and current input with feedback connections. However, a standard RNN cannot handle long-term dependencies that may tend toward an exploding or vanishing gradient. To cope with the problem, Long Short-Term Memory (LSTM) networks were proposed, which bring a “gate” into the standard recurrent cell to improve memory capacity. The inner connections of an LSTM network are shown in Fig. 1. Due to their ability to recall previous information, LSTM networks can capture long-term speech contexts to trace a target speaker. Almost all existing results of RNNs have been performed by LSTM models. Therefore, we hypothesized that LSTM would be more suitable to the current task than the extreme gradient boosting (XGBoost) model, a widely used ensemble model algorithm. The secondary purpose of this study was to compare the performance of the deep learning algorithm with the ensemble model on this task.

Methods

Data Source

We retrospectively analyzed patients whose principal diagnosis was LDH or LSS from June 2013 to June 2020 at Zhongda Hospital, affiliated with Southeast University. Patients who had combined trauma, malignancy, infection, pseudarthrosis, and deformity or incomplete clinical data were excluded. Finally, 1921 patients with free-text admission notes were reviewed, including 1031 patients with LDH and 890 patients with LSS. All diagnoses were confirmed by two senior authors using admission notes, lumbar MRI, and CT scanning. The positive symptoms in the admission notes were reviewed by two researchers.

Data Analysis

Free-text notes were cut by the Jieba package (https://pypi.org/project/jieba/) to build a word set in Python ver-
sion 3.7.6 (https://www.python.org/downloads/release/python-376/). Then all notes were vectorized by the word set. The data set was randomly divided into a training set and testing set at a 7:3 ratio. Two NLP-based machine learning models were developed in this study. LSTM, the most popular RNN, was developed to differentiate LDH and LSS on the training set and was conducted in the Tensorflow version 2.3 keras package (https://tensorflow.google.cn/install), with a dropout approach added for inhibiting overfitting. The LSTM model was compared with the XGBoost model. XGBoost was conducted in Python version 3.7.6 with the scikit-learn package (https://scikit-learn.org/stable/). The NLP algorithms were assessed on the testing set by the following metrics: receiver operating characteristic (ROC) curve, area under the curve (AUC), accuracy score, recall score, F1 score, and precision score.

### Results

In the testing set, the LSTM model achieved an AUC of 0.8487 (Fig. 2), accuracy score of 0.7818, recall score of 0.9045, F1 score of 0.8108, and precision score of 0.7347 (Table 1). In comparison, the XGBoost model achieved an AUC of 0.7565 (Fig. 3), accuracy score of 0.6961, recall score of 0.7387, F1 score of 0.7153, and precision score of 0.6934.

### Discussion

Online medical prediagnosis systems have attracted considerable interest recently as a convenient way to provide professional diagnosis and guidance for patients. Compared with traditional clinical practice, these systems may help avoid unnecessary visits and geographic limitations. Liang et al. applied an AI-based pediatric disease diagnostic system to extract clinically relevant information from EHRs in a manner similar to hypothetico-deductive reasoning. Their model demonstrated high diagnostic accuracies, comparable to experienced physicians, in diagnosing common childhood diseases across multiple organ systems. Zhang et al. presented an AI, traditional Chinese medicine assistive diagnostic system based on unstructured freestyle EHRs, by which 187 commonly known traditional Chinese medicine diseases can be diagnosed and their associated syndromes predicted.

The goal of NLP is to make programs “understand” natural language that, once accessed, exerts the computation ability of AI on processing unstructured big data. AI algorithms have had high repeatability and computation speed in medical tasks that could be used to save clinicians time, allowing them to focus on work requiring more judgment. NLP has been applied for surveillance of adverse events, identification of lesions from radiological reports or images, and prediction of prognosis. However, diagnosis is preliminary to the monitoring of EHRs. To our knowledge, there are no algorithms for automatically diagnosing lumbar disease.

This study does not intend to propose a system of automatic diagnosis based on symptoms; diagnosis must be performed by highly trained surgeons through conscientious physical examination and thoughtful reading of radiological images. In contrast, the analysis could be viewed as an exploratory proof of concept that indicates potential applications of NLP-based machine learning in supporting the most possible diagnosis before formal treatment. The first potential application of current algorithms, such as those in this study, is preliminary screening and diagnosis support in outpatient and referral systems of primary healthcare institutions, which may relieve the pressure caused by the limited medical resources in low-income countries. Another potential application would be as an auxiliary of junior doctors in assessing EHRs. However, we are still a very long way from using NLP automatic diagnosing systems in practical work.

Nonetheless, XGBoost was a great ensemble model algorithm and achieved excellent results in detection of ad-

---

**TABLE 1. Performance of LSTM and XGBoost**

<table>
<thead>
<tr>
<th></th>
<th>LSTM</th>
<th>XGBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy score</td>
<td>0.7818</td>
<td>0.6961</td>
</tr>
<tr>
<td>Recall score</td>
<td>0.9045</td>
<td>0.7387</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.8108</td>
<td>0.7153</td>
</tr>
<tr>
<td>Precision score</td>
<td>0.7347</td>
<td>0.6934</td>
</tr>
</tbody>
</table>
verse events in free-text notes. In previous research, XGBoost showed a lower predictive capability than deep learning on identifying drug use based on Twitter data. Therefore, we hypothesized that LSTM would be more suitable than XGBoost in performing the current task, which was to differentiate LDH and LSS based on positive symptoms. LDH and LSS can have similar symptoms due to a portion of patients with LSS not experiencing typical neurogenic claudication. Without examinations and radiomics, this is hard to differentiate even by surgeons. As a result, the XGBoost model showed underfitting in the internal validation, which reconfirmed the difficulty of this task. However, the LSTM algorithm showed good performance in automatically differentiating LDH and LSS; the AUC and F1 score were greater than 0.8. With cyclic connection, RNNs update the current state based on past states and current input data, which has made RNNs widely adopted in sequential data research, such as text. However, RNNs cannot connect the relevant information when the gap between the relevant input data is large. LSTM can handle the “long-term dependencies” as almost all RNN research has been performed by LSTM models.

Study Limitations

There are several limitations that must be acknowledged. Only candidates for surgery were selected for admission treatment, which inevitably leads to the symptoms of this data set being more serious compared with the overall lumbar degenerative disease population. Wide variability in patients’ expression of symptoms and surgeons’ recordkeeping patterns added another limitation to this work. Furthermore, our current outcomes were restricted to binary results. In future research, a multiple classifier model for diagnosing both more common and rare diseases is needed. Moreover, external validation is also necessary to confirm generalizability of this single-center analysis. Since NLP algorithms are tailored to the specific data with which they were developed, it is not easy to apply them to other EHR systems. The large number of notes revealed the lack of a standardized typology of clinical narratives. Different wording of similar concepts is inevitable due to surgeons being from various regions and having unique note templates and diverse standards of reporting. External validation is a serious challenge for the current LSTM model. Additionally, NLP is prone to overfitting, relating to sample size, missing data, and misclassification. More complete and larger training data sets are necessary to truly reflect patient populations.

Finally, the LSTM model is very deep, which requires large computation and a time-consuming training process. In other words, training an RNN involves significant resources. It will be a hard task to apply and optimize this model with data of larger magnitude. The shortcomings of NLP may signify that the current results, based on single-center data, may be overoptimistic, which probably limits any further application to lumbar diseases.

Conclusions

NLP-based machine learning algorithms were a promising auxiliary to EHRs in spine disease diagnosis. LSTM, the deep learning model, showed better capacity compared with the widely used ensemble model, XGBoost, in differentiation of LDH and LSS using positive symptoms. This study presents a proof of concept for the application of NLP in a prediagnosis system for spine disease.

References


**Disclosures**

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

**Author Contributions**

Conception and design: Ren, Yu. Acquisition of data: P Wang, Zhang. Analysis and interpretation of data: Ren. Drafting the article: Ren, Yu. Critically revising the article: Wu, Xie, Liu, Y Wang. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Wu. Study supervision: Wu, Y Wang.

**Correspondence**

XiaoTao Wu: Zhongda Hospital, Medical College, Southeast University, Nanjing, Jiangsu, China. wuxiaotaospine@seu.edu.cn.