Neurol Ther (2022) 11:1117–1134
https://doi.org/10.1007/s40120-022-00355-7

ORIGINAL RESEARCH

Construction of an Assisted Model Based on Natural Language Processing for Automatic Early Diagnosis of Autoimmune Encephalitis

Yunsong Zhao · Bin Ren · Wenjin Yu · Haijun Zhang · Di Zhao · Junchao Lv · Zhen Xie · Kun Jiang · Lei Shang · Han Yao · Yongyong Xu · Gang Zhao

Received: March 9, 2022 / Accepted: April 7, 2022 / Published online: May 11, 2022

ABSTRACT

Introduction: Early diagnosis and etiological treatment can effectively improve the prognosis of patients with autoimmune encephalitis (AE). However, anti-neuronal antibody tests which provide the definitive diagnosis require time and are not always abnormal. By using natural language processing (NLP) technology, our study proposes an assisted diagnostic method for early clinical diagnosis of AE and compares its sensitivity with that of previously established criteria.

Methods: Our model is based on the text classification model trained by the history of present illness (HPI) in electronic medical records (EMRs) that present a definite pathological diagnosis of AE or infectious encephalitis (IE). The definitive diagnosis of IE was based on the results of traditional etiological examinations. The definitive diagnosis of AE was based on the results of neuronal antibodies, and the diagnostic criteria of definite autoimmune limbic encephalitis proposed by Graus et al. used as the reference standard for antibody-negative AE. First, we automatically recognized and extracted symptoms for all HPI texts in EMRs by training a dataset of 552 cases. Second, four text classification models trained by a dataset of 199 cases were established for differential diagnosis

Yunsong Zhao, Bin Ren and Wenjin Yu have contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40120-022-00355-7.

Y. Zhao · W. Yu · H. Zhang · D. Zhao · J. Lv · G. Zhao (✉)
Department of Neurology, Xijing Hospital, Fourth Military Medical University, Xi'an, China
e-mail: zhaogang@nwu.edu.cn; yun123@fmmu.edu.cn

B. Ren · K. Jiang
Department of Information, Xijing Hospital, Fourth Military Medical University, Xi'an, China

Z. Xie · Y. Xu (✉) · G. Zhao
College of Life Sciences and Medicine, Northwest University, Xi'an, China
e-mail: 20182028@nwu.edu.cn; xuyongy@fmmu.edu.cn

L. Shang
Department of Health Statistics, Fourth Military Medical University, Xi'an, China

H. Yao
Department of Neurobiology, School of Basic Medicine, Fourth Military Medical University, Xi'an, China

of AE and IE based on a post-structuring text dataset of every HPI, which was completed using symptoms in English language after the process of normalization of synonyms. The optimal model was identified by evaluating and comparing the performance of the four models. Finally, combined with three typical symptoms and the results of standard paraclinical tests such as cerebrospinal fluid (CSF), magnetic resonance imaging (MRI), or electroencephalogram (EEG) proposed from Graus criteria, an assisted early diagnostic model for AE was established on the basis of the text classification model with the best performance.

Results: The comparison results for the four models applied to the independent testing dataset showed the naïve Bayesian classifier with bag of words achieved the best performance, with an area under the receiver operating characteristic curve of 0.85, accuracy of 84.5% (95% confidence interval [CI] 74.0–92.0%), sensitivity of 86.7% (95% CI 69.3–96.2%), and specificity of 82.9% (95% CI 67.9–92.8%), respectively. Compared with the diagnostic criteria proposed previously, the early diagnostic sensitivity for possible AE using the assisted diagnostic model based on the independent testing dataset was improved from 73.3% (95% CI 54.1–87.7%) to 86.7% (95% CI 69.3–96.2%).

Conclusions: The assisted diagnostic model could effectively increase the early diagnostic sensitivity for AE compared to previous diagnostic criteria, assist physicians in establishing the diagnosis of AE automatically after inputting the HPI and the results of standard paraclinical tests according to their narrative habits for describing symptoms, avoiding misdiagnosis and allowing for prompt initiation of specific treatment.

Keywords: Autoimmune encephalitis; Computer-assisted diagnosis; Natural language processing; Electronic medical records; Early diagnosis

Key Summary Points

Why carry out this study?

Early diagnosis and etiological treatment can effectively improve the prognosis of patients with autoimmune encephalitis (AE).

There are practical syndrome-based diagnostic criteria for AE described previously.

The sensitivity of these criteria is reasonable but could be further improved.

What was learned from the study?

Using our assisted model, we increased the early diagnostic sensitivity for AE by broadening the number of symptoms analyzed compared to previous criteria.

By using natural language processing technology, the model could provide an early diagnosis of AE automatically, avoiding misdiagnosis and allowing for prompt initiation of specific treatment.

INTRODUCTION

Encephalitis is an inflammatory disease that damages the brain parenchyma and has high morbidity and mortality [1, 2]. According to the etiology, encephalitis can generally be divided into two types: autoimmune encephalitis (AE) and infectious encephalitis (IE). Previous studies have shown that early diagnosis and etiological treatment can effectively improve the prognosis of patients with encephalitis [3–9]. AE refers to a large class of encephalitis associated with anti-neuronal antibodies [10]. At present, there are more than 30 anti-neuronal antibodies related to AE, which can be divided into cell surface antigen antibodies and intracellular antigen antibodies. Among all the antibodies, the most common anti-neuronal antibody is anti-*N*-methyl-D-aspartate receptor (NMDAR) antibody

[11]. Although the clinical presentations of AE are varied, some AE can present with typical manifestations, such as psychiatric symptoms, seizures, short-term memory deficits, decreased level of consciousness, and dyskinesias in anti-NMDAR encephalitis [12].

Currently, a definitive diagnosis of AE is largely dependent on positive test results for anti-neuronal antibodies in cerebrospinal fluid and serum. However, anti-neuronal antibody tests have the following limitations for early diagnosis of AE: (1) many medical institutions do not have the ability to conduct anti-neuronal antibody tests and therefore may delay the optimal treatment time for patients. (2) Some AE cases are negative for anti-neuronal antibodies; thus, even if the results of the anti-neuronal antibody tests are negative, the possibility that encephalitis is immune-mediated cannot be ruled out [12]. The greatest challenge in early diagnosis of AE is differential diagnosis of encephalitis of other etiologies. IE is one of the most common differential diagnoses of AE. Difficulties in the differential diagnosis of AE and IE are mainly reflected in the following aspects: (1) AE often has core symptoms similar to IE; and (2) approximately 50% of patients with AE are negative for anti-neuronal antibodies [13].

To avoid delaying the best opportunity for therapy, Graus et al. [12] proposed criteria for early clinical diagnosis of AE based on typical clinical symptoms and the results of standard paraclinical tests. The clinical symptoms and the standard paraclinical test results, for example, cerebrospinal fluid (CSF), magnetic resonance imaging (MRI), or electroencephalogram (EEG), can be obtained at the early stage of admission. The standard paraclinical tests that are conventional and accessible to most clinicians could be checked as soon as possible. Graus et al. believe that early clinical diagnosis of AE can be obtained and treatment can be carried out quickly using their diagnostic approach. At the same time, the diagnosis and treatment can be adjusted after the results of anti-neuronal antibody testing are clear. As a diagnosis approach, the flowchart proposed from Graus criteria is meant to be used as a whole. One point worth emphasizing is the first criterion for possible autoimmune encephalitis (PAE): the first diagnostic criterion is to screen as many AE cases as possible; thus, high sensitivity is important for PAE criteria to reduce the omission diagnostic rate. Only when the sensitivity of the PAE diagnostic criteria reaches 100% can the subsequent diagnostic process for other subgroups of AE be correctly ensured; otherwise, the omission diagnostic rate may increase [14]. However, the sensitivity of these criteria for PAE has been controversial, ranging from 57.6% to 100.0% [14–17].

The sensitivity of previous diagnostic criteria proposed by Graus et al. [12] is relatively good. On the one hand, the limited number of symptoms (six) that can be incorporated within a clinically usable scale and the simple logic diagnosis rules increase the operability and convenience of clinical practice. On the other hand, the diagnostic criteria may ignore the opportunity of utilizing other symptoms and weighting those symptoms individually. In fact, a great number of atypical initial symptoms can also occur in AE [18–20]. At the same time, there are some negative symptoms in the history of present illness (HPI) in electronic medical records (EMRs) that may aid in differential diagnosis. On the basis of diagnostic criteria proposed previously, taking full advantage of enhancing the initial symptom selection (these atypical positive symptoms and negative symptoms in the HPI described in EMRs) may make some progress in clinical diagnosis of AE via natural language processing (NLP) technology. However, because of the different habits of physicians in writing EMRs, there are a variety of synonyms for the same symptom in raw texts. If we want to analyze the symptoms, we should recognize and normalize the symptoms that have synonyms.

For a long time, comprehensive analysis of the characteristics of AE symptoms through traditional labor-dependent methods has represented a great challenge because the symptoms are complex and diverse. NLP is a subfield of linguistics, computer science, and artificial intelligence. The technology can accurately process, extract, and analyze large amounts of information from natural language data. Clinical named entity recognition (CNER) [21] and

text classification [22] are both common missions in NLP technology, which could be applied to the healthcare domain for automatically extracting complex and diverse clinical symptoms or classifying the clinical texts from a large corpus of EMRs, using smaller subset of cases for training models. The bidirectional long short-term memory conditional random field (BiLSTM-CRF) model is one of the most popular models to achieve the CNER mission [23, 24]. Equally important, following the arrival and development of the bidirectional encoder representations from transformers (BERT) [25, 26], CNER has started to adopt the BERT-based approach, which brings the advantage of allowing pretrained models [27]. Text classification usually needs to achieve feature selection at first. Feature selection is the process of selecting the features which contribute most to the classification of a given text. Text classification methods and feature selection methods are rapidly growing [28]. To conclude, with the rapid development of NLP technology [29–31], we may have the opportunity to recognize and extract medical terminology, especially for symptoms described in the HPI in EMRs; perform a comprehensive analysis of clinical symptoms; and develop an assisted diagnosis method for PAE. Our study aims to develop an assisted model for early clinical diagnosis of AE based on NLP technology and to compare its sensitivity with that of previously established criteria from Graus et al. [12].

METHODS

Study Population and Design

Our data included 2514 Chinese EMRs with a diagnosis of central nervous system (CNS) infectious or inflammatory diseases, which were identified using the International Classification of Diseases-Tenth Revision (ICD-10) codes from the medical records database of the neurology department of Xijing Hospital during a period of 10 years (October 2010–December 2020). The ICD-10 codes are provided in Table S1 in the supplementary material. De-identification of all EMRs was conducted for every known identifier type (ID, name, gender, age, address, etc.). Standard paraclinical tests (e.g., CSF, MRI, or EEG studies) and other routine laboratory data were obtained by review of the EMRs. The Xijing Hospital Ethics Committee approved this study (KY20192071-F-1).

Our study constructed two datasets by using 2514 Chinese EMRs: the first to train CNER as part of the NLP (552 cases) analysis pipeline, and the second (199 cases) to train and test the diagnosis classification model between AE and IE. The two datasets were filtered respectively from 2514 cases, so there was overlap between the CNER dataset and the text classification model dataset. To eliminate subjective interference and evaluate the classification model more objectively, the EMR samples included in the text classification dataset were AE or IE cases with definite evidence of etiology. The definitive diagnosis of IE was based on the results of traditional etiological examinations, which included microscopic staining, pathogenic microbiological analysis, and PCR. The definitive diagnosis of AE was based on the exclusion of other definite causes (e.g., IE). Then, all patients underwent an extensive search for neuronal antibodies and they required positive antibodies, either in CSF or serum, using commercial cell-based assay kits, according to published guidelines [32, 33]. Finally, because the possibility that autoantibodies may not be detected in definite autoimmune limbic encephalitis, the clinical diagnostic criteria proposed by Graus et al. only applied to definite autoimmune limbic encephalitis in this study [12].

A synopsis of the overall NLP analysis pipeline is shown in Fig. 1. First, we automatically extracted symptoms (i.e., CNER) from all HPI texts by training the BiLSTM-CRF model on the basis of a dataset of 552 cases with a single diagnosis of central nervous system (CNS) infection or AE. Second, post-structuring of the HPI in EMRs for all cases was implemented after normalizing symptom terminologies in English language. Third, four text classification models trained by a dataset of 199 cases were established for differential diagnosis of AE and IE based on a post-structuring text dataset of every HPI. The optimal model was identified by

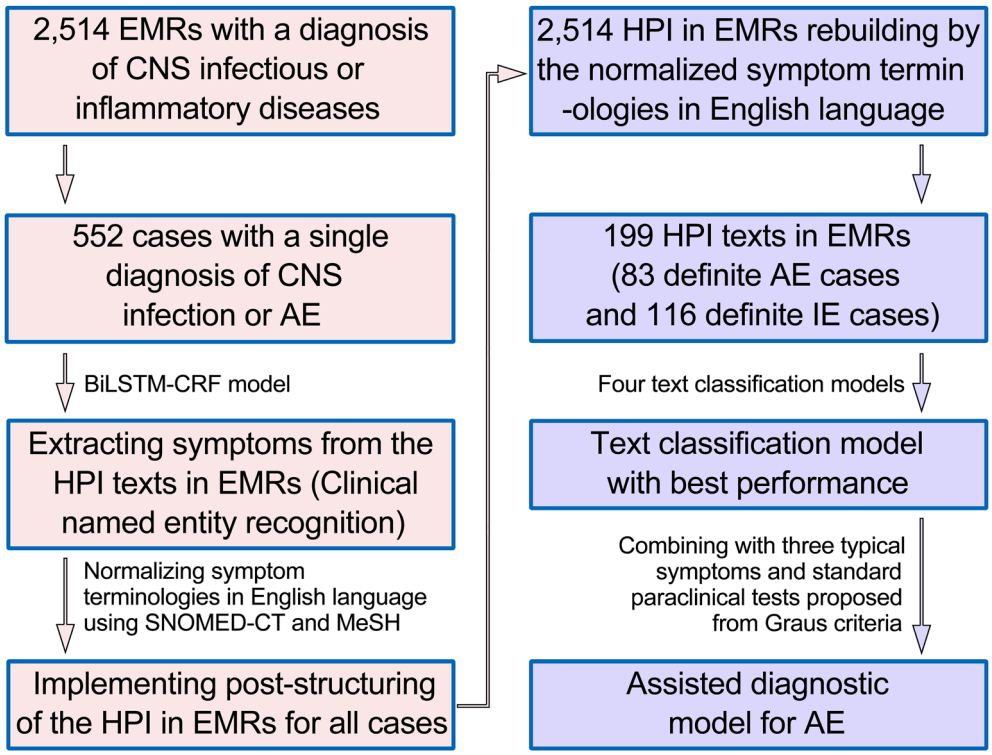

Fig. 1 Synopsis of the overall NLP analysis pipeline. *NLP* natural language processing, *EMRs* electronic medical records, *CNS* central nervous system, *AE* autoimmune encephalitis, *BiLSTM-CRF* bidirectional long short-term memory conditional random field, *HPI* history of present illness, *SNOMED-CT* Systematized Nomenclature of Medicine-Clinical Terms, *MeSH* Medical Subject Headings, *IE* infectious encephalitis

evaluating and comparing the performance of the four models. Finally, combined with three typical symptoms and the results of standard paraclinical tests (e.g., CSF, MRI, or EEG studies) proposed from Graus criteria, an assisted early diagnostic model for AE was established on the basis of the text classification model with the best performance.

Data Preprocessing for CNER

We filtered all 552 patient records identified with a single diagnosis of CNS infection or AE from 2514 EMRs with CNS infectious or inflammatory diseases. While all 552 cases were used at different stages of the CNER development (training word embedding, identifying the symptoms), a random subset of 140 cases (25% of 552 patient records) were selected for manual word annotations to assist with training. It has been proven that Chinese medical text segmentation is very important for producing high-quality word embedding and promoting downstream information extraction applications [34]. Therefore, we used the Jieba Chinese Word Segmentation Library supported by Python programming language to segment the HPI in EMRs.

In our CNER approach, annotated data were represented in the BMESO format, in which each word was assigned to one of five classes: B, beginning of an entity; M, middle of an entity; E, ending of an entity; S, single word for an entity; O, outside of an entity. Therefore, the CNER problem became a classification problem requiring assignment of one of the five class tags to each word. The annotation guidelines were similar to those in Yang et al.'s study [35]. One main difference was that we only manually annotated symptoms in the HPI in EMRs. Thus,

we only had one type of entity in this study. Another difference was that negative symptoms were recognized as a whole entity [36]. The statistics of the training subset of the HPI used for CNER is shown in Table S2 in the supplementary material. There were 26,655 words and 1055 symptoms in the HPI in the training subset, and B, M, E, S, and O were annotated 1006, 479, 1006, 49, and 24,115 times, respectively.

We explained the annotation guidelines in this study to three specialized neurology physicians. The final annotation results were identified by the following rules: Word boundaries of symptoms were marked using B, M, E, S, and O tags by manual annotations performed by two physicians. Examples of some annotated sequences are provided in Table S3 in the supplementary material. After the manual annotation by the two physicians was completed, the inter-annotator agreement was calculated with Kohen's kappa and was 0.87. Then, the manual annotation results provided by the two physicians were compared. When the results were the same, the same annotation result was the consensus result. In the case of different annotation results by the two physicians, the third physician made a final interpretation of the two annotation results, and one was selected as the consensus result.

Training and Evaluation for CNER Using the BiLSTM-CRF Model

To improve the quality of training word embedding, 552 HPI texts were used as a corpus for training word embedding. We used 140 HPI texts for the BiLSTM-CRF model. The Word2Vec tool is used to train word embedding using a continuous bag of words (CBOW) model or skip-gram model [37]. Various empirical evidence shows that the CBOW model performs better than the skip-gram model for databases with only hundreds of thousands of words [38, 39]. Therefore, the CBOW model was adopted in this study to achieve word embedding. The dimension of the word embedding vector was set to 128 dimensions, and the other hyper-parameters are provided in Table S4 in the supplementary material.

The dataset, which contained 140 HPI texts, was split into two mutually exclusive subsets: training (80% of dataset) and testing (20% of dataset) subsets [40]. We divided every HPI with full stops into sentences. Each sentence of every HPI was used as an input sequence in the following model. The implementation of CNER in this study is based on the BiLSTM-CRF model, with word embedding as the input of sequences [23, 24]. To optimize the hyper-parameters of the BiLSTM-CRF model, a tenfold cross-validation method was adopted using the training subset. When the optimal hyper-parameters of the BiLSTM-CRF model were determined, the entire training subset was trained as the final BiLSTM-CRF model according to the optimal hyper-parameters.

The hyper-parameters of the BiLSTM-CRF model were fine-tuned, training the model with the training subset and evaluating it with the *F* measure. The results for each configuration of the parameters (batch size, number of epochs, dropout, and learning rate with the Adam algorithm [41]) in the hyper-parameter fine-tuning stage are shown in Fig. S1 in the supplementary material. In the case of the embedding created from words by Word2Vec, the best hyper-parameters were as follow: 2 batches, 40 epochs, 0.2 dropout, and 0.001 learning rate. The other hyper-parameters are provided in Table S4 in the supplementary material.

The performance of the final BiLSTM-CRF model was measured by precision, recall, and *F* measure for all entities using the independent testing subset [42]. The evaluation program provided two sets of measures—exact match and inexact match—where exact match means that an entity is correctly predicted if, and only if, the starting and ending offsets are exactly the same as those in the consensus result; the inexact match means that an entity is correctly predicted if it overlaps with any entity in the consensus result [43, 44]. Table S5 in the supplementary material shows the statistics of the independent testing subset of the HPI used in this study. There were 6425 words and 296 symptoms in the HPI in EMRs. Table S5 also shows the performance of the BiLSTM-CRF model applied to the independent testing subset. The numbers in columns 4–6 are precision,

recall, and *F* measure values for all entities using the exact match or inexact match measures.

Post-Structuring of the HPI

The final BiLSTM-CRF model was used to identify and extract the symptoms in 552 HPI texts presented in EMRs. All the symptoms were classified into different groups according to whether they had synonyms. Then, we obtained normalized symptom terminologies by establishing a mapping relationship between the categorized symptoms and the international standard English medical terms set, Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and Medical Subject Headings (MeSH). The process of normalization of symptoms was conducted using Python regular expressions according to normalized symptom terminologies in English language. Then, the post-structuring of the HPI was completed using the normalized symptom terminologies; in other words, we rebuilt the HPI texts using the normalized symptom terminologies, which included typical symptoms, atypical symptoms, and negative symptoms.

Training and Evaluation for Text Classification Model

According to the definite diagnostic criteria for AE and IE, we obtained the text classification dataset of 199 cases by reviewing EMRs. The dataset included 83 AE cases and 116 IE cases. There are several autoimmune CNS diseases (primary CNS angiitis, Rasmussen's encephalitis) that are often considered in the differential diagnosis of autoimmune encephalitis because of the clinical features. These diseases were considered immune-related diseases and excluded [11, 12]. The overall data were randomly divided into a 65% training dataset and a 35% testing dataset. The training dataset included 53 AE cases and 75 IE cases, and the independent testing dataset included 30 AE cases and 41 IE cases.

We did not directly perform text classification on raw texts because of the impact of the EMR template. The EMR template makes different EMRs have the same words, which are not related to symptoms of AE. However, the large numbers of same words from EMR template would cover up the role of symptoms of AE and affect the result of text classification. Meanwhile, as a result of different habits in writing EMRs, there were a variety of synonyms for the same symptom in raw texts. Therefore, our text classification model was based on rebuilding the HPI using normalized symptom terminologies rather than the raw HPI texts presented in EMRs. Our text classification models consisted of four models: two different classifiers with two different text feature selection methods respectively. Four text classification models, namely, naïve Bayesian classifier (NBC) and support vector machine (SVM), using two different text feature selection methods, namely, bag of words (BoW) and term frequency–inverse document frequency (TF-IDF), were established to distinguish AE and IE on the basis of structured texts of the HPI, which were rebuilt using normalized symptom terminologies. The process of determining the optimal hyper-parameters of the four text classification models was realized by the grid search function GridSearchCV from the Scikit-Learn library [45]. The hyper-parameters are provided in Table S4 in the supplementary material.

When the optimal hyper-parameters of the text classification models were determined, the final models were trained using the whole training dataset. The performances of the four text classification models were evaluated and compared using the independent testing dataset. The performances of the text classification models were measured by sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUROC). Finally, the model that performed best was used to establish the assisted diagnostic model for AE.

Evaluation of the Assisted Diagnostic Model

Because the diagnostic criteria from Graus et al. [12] emphasized the importance of psychiatric symptoms, seizures, and short-term memory deficits, logical diagnosis rules for these three

symptoms were added to the basis of the text classification model. Specifically, when the text classification model judges that the case is not an AE case, if the case contains one or more of these three symptoms, the case will be clinically diagnosed as an AE case. Furthermore, our assisted diagnostic model was combined with the results of standard paraclinical tests (e.g., CSF, MRI, or EEG studies) from the Graus criteria.

All cases of the independent testing dataset were analyzed according to the diagnostic criteria for PAE from Graus et al. and the assisted diagnostic model to compare the etiological diagnosis. The performance was measured and assessed according to sensitivity, specificity, accuracy, and confusion matrices.

We analyzed demographics, clinical, and standard paraclinical test characteristics. Continuous variables were presented as the mean ± standard deviation (SD) in the descriptive analyses, while categorical and binary variables were presented as frequencies (n) and percentages (%). Student's t test and chi-squared test were used to compare outcomes between patient subgroups for continuous and categorical data, respectively. All data acquisition, processing, and analyses were conducted in the Python programming language (version 3.7.0) [46, 47], TensorFlow library [48, 49], and Scikit-Learn library [45].

RESULTS

Our text classification dataset included 199 cases (83 definite AE cases and 116 definite IE cases). The process of constructing the text classification dataset is shown in Fig. 2. Demographics characteristics and standard paraclinical test characteristics are compared between AE and IE groups in Table 1. Compared to the IE group, the AE group had a significantly higher rate of cases with EEG abnormalities (57.8% vs. 25.9%; $P < 0.001$), and a lower percentage of male cases (44.6% vs. 62.1%; $P = 0.015$). There was no significant between-group difference with respect to age, CSF pleocytosis (white blood cell count greater than 5/mm^3), CSF-

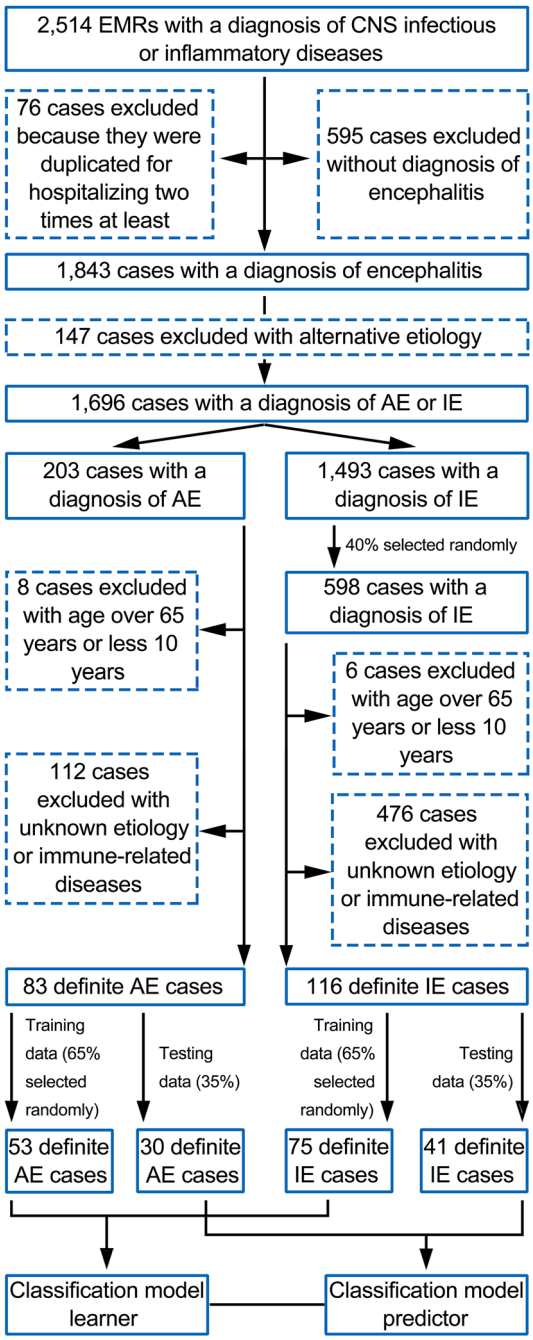

Fig. 2 Flowchart illustrating the development of an assisted model based on NLP for the diagnosis of AE. *NLP* natural language processing, *EMRs* electronic medical records, *CNS* central nervous system, *AE* autoimmune encephalitis, *IE* infectious encephalitis

Table 1 Comparison of demographics characteristics and standard paraclinical test characteristics between AE and IE groups (n = 199)

Variable	AE (n = 83)	IE (n = 116)	P value[a]
Demographic characteristics			
Age (years)	31 (34.7 ± 15.5)	37 (36.0 ± 13.7)	0.533
Sex, male	37 (44.6)	72 (62.1)	0.015*
Standard paraclinical test characteristics			
CSF pleocytosis (white blood cell count > 5/mm^3)	53 (63.9)	87 (75.0)	0.09
CSF-specific oligoclonal bands positive	2 (2.4)	1 (0.9)	0.769
MRI abnormalities	40 (48.2)	48 (41.4)	0.340
EEG abnormalities	48 (57.8)	30 (25.9)	0.000**

Data are presented as the mean ± standard deviation (SD) or as n (%)

AE autoimmune encephalitis, *IE* infectious encephalitis, *CSF* cerebrospinal fluid, *MRI* magnetic resonance imaging, *EEG* electroencephalogram

[a]Significant difference at $^*P < 0.05$ and $^{**}P < 0.001$

specific oligoclonal bands positive, or MRI abnormalities.

By using the BiLSTM-CRF model, we recognized and extracted 5819 symptoms from 552 HPI in EMRs identified with the single diagnosis of CNS infection or AE. After normalizing synonyms of symptom terminologies using SNOMED-CT and MeSH, we finally obtained 219 normalized symptom terminologies in English language. The 219 normalized symptom terminologies and the proportion of the IE and AE cohorts with each symptom are shown in Table S6 in the supplementary material. In the definite AE group, the majority of patients presented with psychiatric symptoms (74.7%), decreased level of consciousness (68.7%), seizures (60.2%), and involuntary movement (59.0%). In the IE group, headache (81.9%) and fever (73.3%) were frequently encountered. Significant differences (only those with $P < 0.001$) with the frequency of clinical characteristics for comparing between IE and AE groups are shown in Fig. 3. The autoantibodies and microorganisms detected in samples from definite AE and IE cases are shown in Fig. 4.

Comparing the results of accuracy, sensitivity, and specificity determined using the independent testing dataset in Fig. 5 shows that NBC with the BoW model performs better than the other three models on the basis of accuracy and sensitivity, obtaining 84.5% (95% confidence interval [CI] 74.0–92.0%) and 86.7% (95% CI 69.3–96.2%), respectively. The AUROC comparison results for the four models applied to the independent testing dataset are shown in Fig. 6, which shows that the AUROC of NBC with BoW achieved the best performance, with an AUROC of 0.85. However, the examination of specificity metrics suggested that the NBC with TF-IDF model, SVM with BoW model, and SVM with TF-IDF model achieved 85.4% (95% CI 70.8–94.4%), which is better than the 82.9% (95% CI 67.9–92.8%) observed with the NBC with BoW model. Since the goal of our study was to improve the sensitivity of the clinical diagnosis of AE, the NBC with BoW model was selected as the best model after comprehensive consideration of various performance measures.

We used the sensitivity, specificity, and accuracy to evaluate the diagnostic performance for PAE between the assisted diagnostic model and the diagnostic criteria from Graus et al. using the same independent subset of patients (30 definite AE cases and 41 definite IE cases). The performance measures show that our model achieved better sensitivity, specificity, and accuracy than the diagnostic criteria

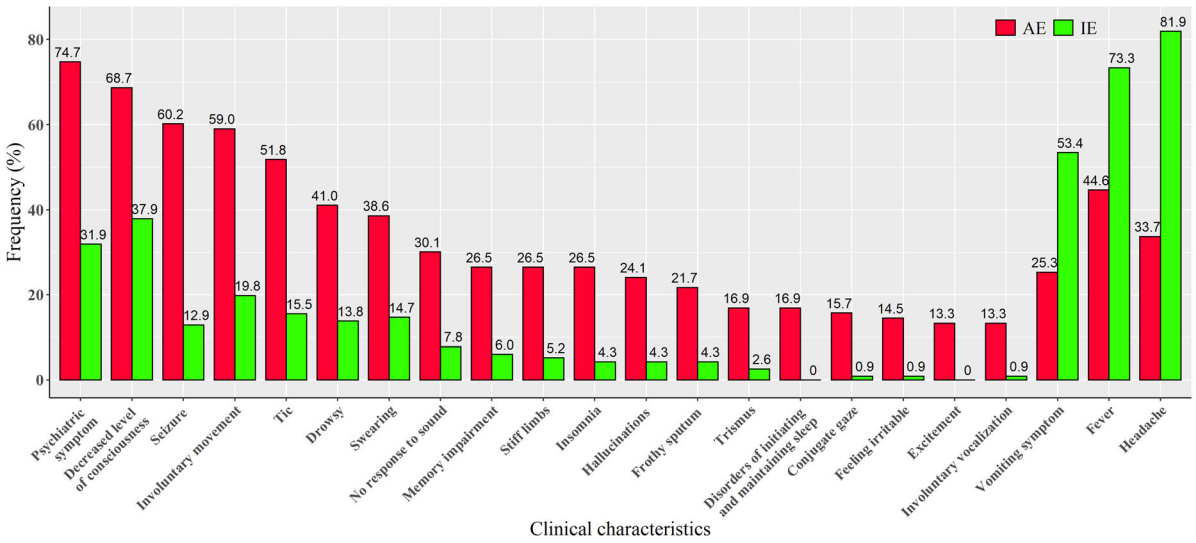

Fig. 3 Frequency of clinical characteristics for comparing between IE and AE groups (significant difference at $P < 0.001$). *AE* autoimmune encephalitis, *IE* infectious encephalitis

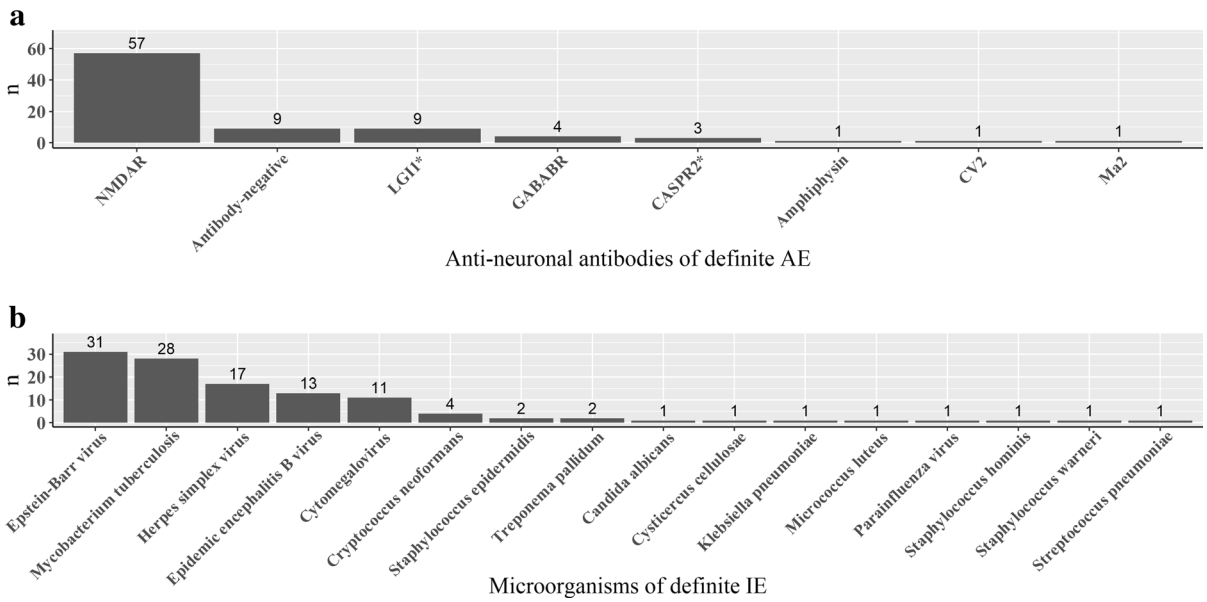

Fig. 4 Autoantibodies and microorganisms detected in samples from definite AE and IE of the whole dataset. **a** Antibodies detected in samples from definite AE of the whole dataset ($n = 83$); asterisk indicates the antibodies were found in the same patient. **b** Microorganisms detected in samples from definite IE of the whole dataset ($n = 116$). *AE* autoimmune encephalitis, *IE* infectious encephalitis, *NMDAR* anti-*N*-methyl-D-aspartate receptor, *LGI1* anti-leucine-rich glioma-inactivated protein 1, $GABA_BR$ anti-gamma-aminobutyric acid B receptor, *CASPR2* anti-contactin-associated protein 2

proposed by Graus et al. The performance is as follow: 86.7% (95% CI 69.3–96.2%) vs. 73.3% (95% CI 54.1–87.7%), 75.6% (95% CI 59.7–87.6%) vs. 56.1% (95% CI 39.7–71.5%), and 80.3% (95% CI 69.1–88.8%) vs. 63.4% (95% CI 51.1–74.5%), respectively. The detailed

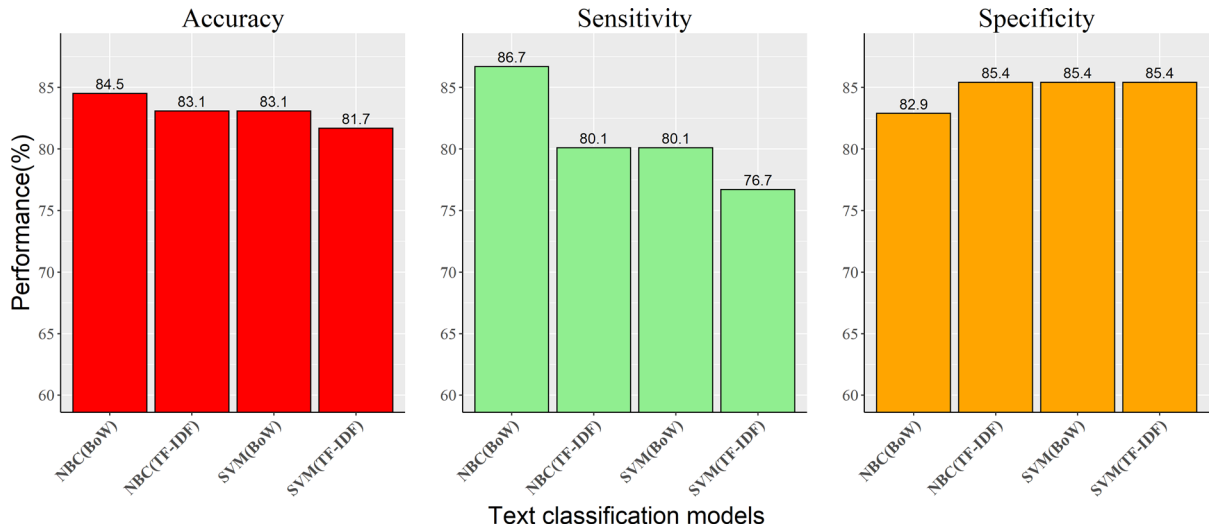

Fig. 5 Predictive performance of four text classification models to distinguish AE and IE on the basis of symptoms of the HPI using the independent testing dataset. *AE* autoimmune encephalitis, *IE* infectious encephalitis, *HPI* history of present illness, *NBC* naïve Bayesian classifier, *SVM* support vector machine, *BoW* bag of words, *TF-IDF* term frequency–inverse document frequency

results for the numbers of true positive, false positive, true negative, and false negative are shown in confusion matrices (Fig. 7).

DISCUSSION

In this study, we have developed an assisted diagnostic model for AE based on NLP technology by recognizing symptoms described in EMRs using CNER. We subsequently tested this model in an independent testing dataset and compared its performance with previously established Graus criteria. Compared with the diagnostic criteria proposed previously, the early diagnostic sensitivity for PAE using the assisted diagnostic model based on the independent testing dataset was improved from 73.3% (95% CI 54.1–87.7%) to 86.7% (95% CI 69.3–96.2%). Since IE is the most common and most difficult disease to differentiate from AE in clinical work, IE is taken as the most important disease for differential diagnosis of AE. In the past, early clinical diagnosis of AE was mostly based on the diagnostic criteria proposed by Graus et al. in 2016, but the consistency of the performance of the diagnostic criteria for PAE is controversial, ranging from 57.6% to 100.0% [14–17]. Our study presented some of the results of previous studies that evaluated and validated the performance of Graus criteria and compared them with our study (Table 2). The differences in sensitivity for PAE in previous studies may be caused by the following reasons: (1) sample size of cases is small and different. (2) Definition of the AE group is different, e.g., Wagner et al.'s AE group contained 17 definite AE and 16 probable AE cases. However, the sensitivity for PAE would be 29.4% if they excluded cases which diagnosed as probable AE, only bringing the 17 definite AE cases into the AE group. (3) Ratio of different anti-neuronal antibody cases varies, such as a higher ratio of NMDAR cases in the AE group, for which the algorithm seems to have a particularly high sensitivity. (4) Sensitivity for PAE increases with time; the sensitivity reported by Li et al. for the time period of up to 14 days after admission only was 60.4% for PAE.

In our study, the sensitivity increased from 73.3% to 86.7% by using the assisted diagnostic model, which compared with the Graus diagnostic criteria applied to the same independent testing dataset. We reviewed the reasons for misdiagnosis of the four cases in the independent testing dataset when using the assisted diagnostic model. Two were due to no

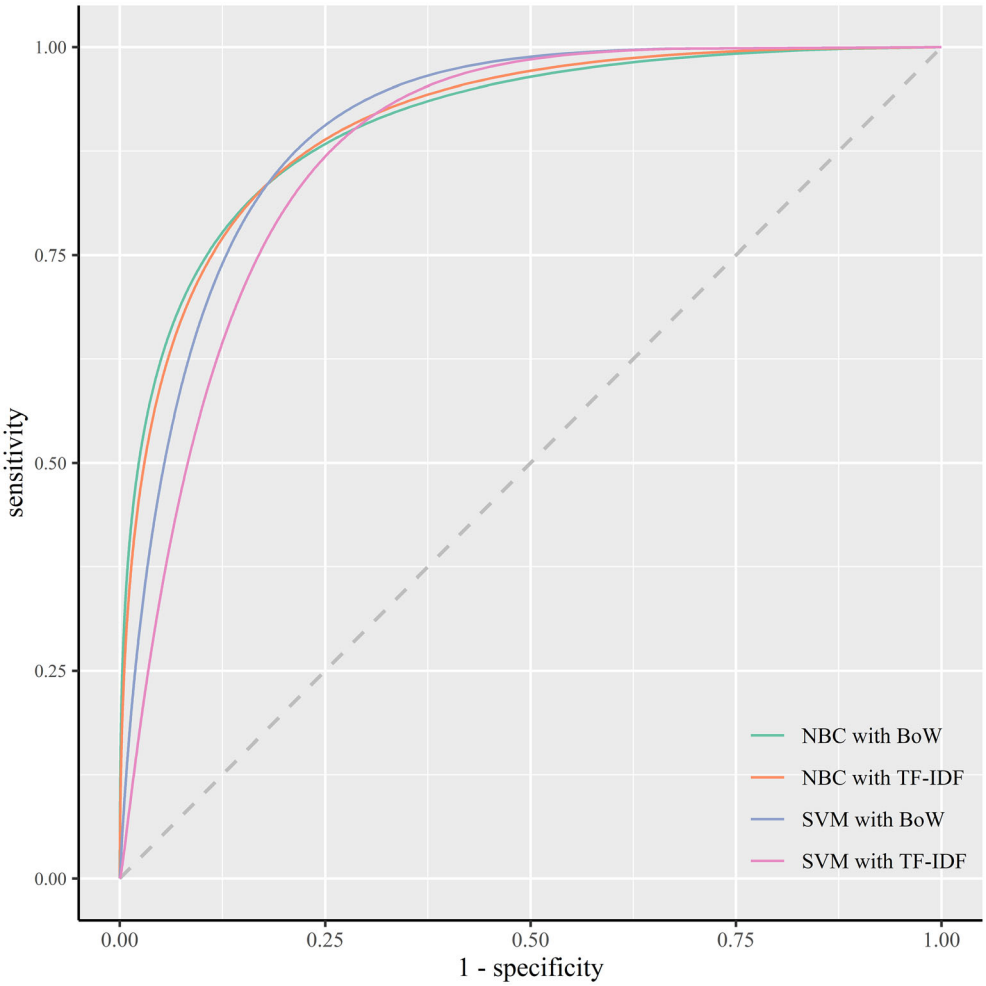

Fig. 6 AUROC of text classification models distinguished AE from IE using independent testing dataset. NBC with BoW, 0.85; NBC with TF-IDF, 0.83; support vector machine (SVM) with BoW, 0.83; SVM with TF-IDF, 0.81. *AUROC* area under the receiver operating characteristic curve, *AE* autoimmune encephalitis, *IE* infectious encephalitis, *NBC* naïve Bayesian classifier, *SVM* support vector machine, *BoW* bag of words, *TF-IDF* term frequency–inverse document frequency

abnormalities in the CSF, MRI or EEG, and the other two were due to atypical symptoms, which were not identified by the assisted diagnostic model. However, our model identified four more true positive AE cases than the diagnostic criteria for PAE from Graus et al.

The measure of specificity mainly depends on the type of disease that is used for the differential diagnosis of AE. If the disease is IE, differential diagnosis is difficult, leading to relatively poor specificity of PAE both for Graus criteria and our assisted model. If the disease is a non-AE disease that can easily be identified, then specificity would be relatively high. In fact, compared with typical symptoms associated with AE and the results of standard paraclinical tests, "Reasonable exclusion of alternative causes" in diagnostic criteria for AE is the most important factor for enhancing the specificity [17]. Therefore, the specificity requires much evidence to exclude all other etiologies, and our study aims to improve the AE diagnostic sensitivity, (finding out the highest number of cases of PAE), not the diagnostic specificity.

In the early differential diagnosis of AE, IE is the most common and difficult disease to

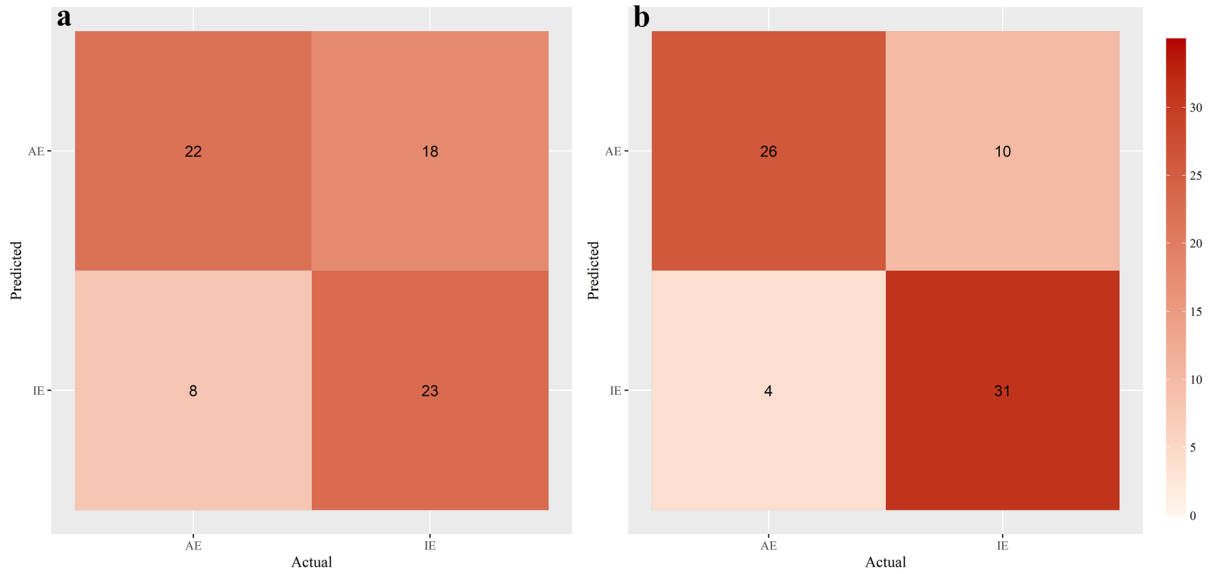

Fig. 7 Confusion matrices of different diagnostic methods for AE using the independent testing dataset. **a** Confusion matrix for predictions of Graus criteria vs. etiology diagnosis of AE; **b** Confusion matrix for predictions of assisted diagnostic model vs. etiology diagnosis of AE. *AE* autoimmune encephalitis, *IE* infectious encephalitis

differentiate, especially viral encephalitis [50]. In our study, the assisted diagnostic model misdiagnosed 10 IE cases as AE, of which eight cases were viral encephalitis and the other two cases were encephalitis caused by *Mycobacterium tuberculosis*. Since infection itself is an important cause of AE, an increasing number of studies have begun to focus on AE caused by viral encephalitis [19, 51]. This infection-induced AE is not only difficult to diagnose but also may induce double-peak encephalitis, which brings great challenges in clinical treatment. For physicians, fundamentally solving the differential diagnosis problem between AE and IE requires an accurate diagnosis, namely, an etiological diagnosis. Early clinical diagnosis of AE can only be adopted as a compromise measure to allow for early initial treatment until an etiological diagnosis can be provided. A proactive search for the exact etiology of AE or IE should always be carried out through the clinical diagnosis and treatment process.

In our study, all the possible accompanying symptoms of AE (including typical positive symptoms, atypical positive symptoms and negative symptoms) were considered in the development of the assisted diagnosis model. After extracting symptoms from the HPI texts, we normalized synonyms of symptom terminologies in English language using SNOMED-CT and MeSH. Then post-structuring of the HPI texts and development of the text classification model were implemented using symptoms terminologies in English language. This facilitates generalization of our findings and our assisted diagnostic model to English speaking populations. The above method not only accounted for the clinical characteristics of complex and diverse AE symptoms but also effectively improved the sensitivity and specificity of early clinical diagnosis of AE. Physicians could quickly obtain an early clinical diagnosis of AE when they input the HPI into the model according to their narrative habits for describing symptoms in Chinese language or English language for non-Chinese speakers.

There are several limitations to this study: First, the Graus diagnostic criteria were not developed to differentiate AE from IE, but to allow for early diagnostic suspicion of AE among all patients with neurologic syndromes. This is a limitation that we used Graus criteria in comparison with our model to distinguish AE from IE. Our assisted diagnostic model shows

Table 2 Summary of early diagnostic performance in previous studies using Graus criteria and comparison with our study

Studies	Diagnostic methods for PAE	AE (*n*)	IE or non-AE[a] (*n*)	Sensitivity (%)	Specificity (%)
Giordano et al. [16]	Graus criteria	22	0	100.0	–
Baumgartner et al. [15]	Graus criteria	50	0	86.0	–
Li et al. [14]	Graus criteria	64	31[a]	84.3	93.5
Wagner et al. [17]	Graus criteria	33	51	57.6	7.8
Our study	Graus criteria	30	41	73.3	56.1
Our study	Assisted model	30	41	86.7	75.6

AE autoimmune encephalitis, *IE* infectious encephalitis, *PAE* possible autoimmune encephalitis
[a]Cases with non-autoimmune encephalitis in Li et al.'s study, including viral encephalitis, purulent encephalitis, tuberculous meningoencephalitis, central nervous system tumor, and epileptic disorders

performance only for the differential diagnosis between AE and IE, which would limit generalizability of our study. There are other differential diagnoses can also be challenging in the evaluation of PAE except IE (e.g., new onset epilepsy due to other causes, other organic encephalopathies, psychiatric disorders, etc.). In addition, patients with antibody-associated AE may have different clinical features to those with antibody-negative AE. However, the antibody-negative AE cases in our study only included the ones with definite limbic encephalitis, which further limits generalizability. In the future, we plan to include more diseases, which need differential diagnosis with AE, and analyze texts in EMRs from patients in primary care, general medical, or psychiatric settings because the greatest unmet need with regard to early diagnosis arises from patients not yet evaluated in neurology departments.

Second, NLP research is currently dominated by the use of transformer models [52], such as BERT. In the future, we plan to adopt advanced BERT-based transformer models (e.g., RoBERTa [53], ELECTRA [54]) or specialized models (e.g., MacBERT [55] and BioBERT [56]) as pretrained models, which could be applicable in the present work.

Third, model explainability was not discussed in our study. However, the need for explainability is present in NLP [57] and the healthcare domain in particular, which is currently a very active area of research [58–60]. The complexity arising from the large parameter space and the combination of algorithms makes models uninterpretable for humans, i.e., the decision process cannot be fully comprehended [61, 62]. To alleviate the issues present in explainability, an assistive diagnostic model that humans can understand, manage, and trust should be proposed in our future work [63].

Finally, our study is retrospective. The sample size of datasets is small, which includes 83 definite AE cases and 116 definite IE cases. As a result of the small sample size, deep neural networks were not used to perform text classification. To explore the effectiveness of the assisted diagnostic model for early clinical diagnosis of AE, larger prospective studies should be conducted to obtain more powerful clinical evidence.

CONCLUSION

In this study, we described the development of an assisted diagnostic model for AE based on normalized symptoms from the HPI described in EMRs via NLP technology. We demonstrated that the assisted diagnostic model could effectively increase diagnostic sensitivity for AE compared to previous diagnostic criteria from Graus et al. This model is capable of assisting physicians in establishing the diagnosis of AE automatically after inputting the HPI and the results of standard paraclinical tests according to their narrative habits for describing

symptoms, allowing for more accurate diagnosis and prompt initiation of specific treatment.

ACKNOWLEDGEMENTS

We thank the scientific secretaries Rui Wu and Huimin Zhou for their invaluable help in preparing the source data that supported this manuscript.

Funding. This work was supported by the National Natural Science Foundation of China (grant number 81671185). The Rapid Service Fee was funded by the corresponding author, Gang Zhao.

Editorial Assistance. We would like to thank AJE. (https://www.aje.com/) for English language editing. Gang Zhao, the corresponding author, provided funding for the English language editing assistance from AJE.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Author Contributions. All authors contributed to the study conception and design. Yunsong Zhao, Bin Ren and Wenjin Yu were the primary authors of the article. Yunsong Zhao, Bin Ren and Lei Shang contributed in data analyses. Yunsong Zhao, Wenjin Yu, Kun Jiang and Han Yao contributed data acquisition, data interpretation and manuscript preparation. Haijun Zhang, Di Zhao, Junchao Lv and Zhen Xie contributed significantly to the text annotation and review of the electronic medical records. Yongyong Xu and Gang Zhao made substantial contributions to the conception and design of the study and revised the article critically for important intellectual content. All authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

Disclosures. Yunsong Zhao, Bin Ren, Wenjin Yu, Haijun Zhang, Di Zhao, Junchao Lv, Zhen Xie, Kun Jiang, Lei Shang, Han Yao, Yongyong Xu and Gang Zhao have nothing to disclose.

Compliance with Ethics Guidelines. The studies involving human participants were reviewed and approved by Xijing Hospital Ethics Committee (KY20192071-F-1). The patients/participants provided their written informed consent of cerebrospinal fluid routine examination before lumbar puncture at the time this was conducted.

Data Availability. All data generated or analyzed during this study are included in this published article/as supplementary information files.

REFERENCES

1. Pewter SM, Williams WH, Haslam C, Kay JM. Neuropsychological and psychiatric profiles in acute encephalitis in adults. Neuropsychol Rehabil. 2007;17(4–5):478–505. https://doi.org/10.1080/09602010701202238.

2. Zhao L, Zhou M, Wang B, Guo J, Chen N, He L. Clinical characteristics and outcome of clinically diagnosed viral encephalitis in southwest China. Neurol Sci. 2015;36(12):2191–7.

3. Titulaer MJ, McCracken L, Gabilondo I, et al. Treatment and prognostic factors for long-term outcome in patients with anti-NMDA receptor encephalitis: an observational cohort study. Lancet Neurol. 2013;12(2):157–65. https://doi.org/10.1016/S1474-4422(12)70310-1.

4. Nosadini M, Eyre M, Molteni E, et al. Use and safety of immunotherapeutic management of *N*-methyl-D-aspartate receptor antibody encephalitis: a meta-analysis. JAMA Neurol. 2021;78(11):1333–44. https://doi.org/10.1001/jamaneurol.2021.3188.

5. Heth JA. Neurosurgical aspects of central nervous system infections. Neuroimaging Clin N Am. 2012;22(4):791–9.

6. Dorsett M, Liang SY. Diagnosis and treatment of central nervous system infections in the emergency department. Emerg Med Clin N Am. 2016;34(4):917–42. https://doi.org/10.1016/j.emc.2016.06.013.

7. Poissy J, Wolff M, Dewilde A, et al. Factors associated with delay to acyclovir administration in 184 patients with herpes simplex virus encephalitis. Clin Microbiol Infect. 2009;15(6):560–4. https://doi.org/10.1111/j.1469-0691.2009.02735.x.

8. Wang W, Li J-M, Hu F-Y, et al. Anti-NMDA receptor encephalitis: clinical characteristics, predictors of outcome and the knowledge gap in southwest China. Eur J Neurol. 2016;23(3):621–9.

9. Peng Y, Liu X, Pan S, Xie Z, Wang H. Anti-*N*-methyl-D-aspartate receptor encephalitis associated with intracranial *Angiostrongylus cantonensis* infection: a case report. Neurol Sci. 2017;38(4):703–6.

10. Hernández Ramos FJ, Palomino García A, Jiménez Hernández MD. Antibody-mediated encephalitis. Med Clin (Barc). 2021;156(6):302–4.

11. Yu Y, Wu Y, Cao X, et al. The clinical features and prognosis of anti-NMDAR encephalitis depends on blood brain barrier integrity. Mult Scler Relat Disord. 2021;47: 102604. https://doi.org/10.1016/j.msard.2020.102604.

12. Graus F, Titulaer MJ, Balu R, et al. A clinical approach to diagnosis of autoimmune encephalitis. Lancet Neurol. 2016;15(4):391–404. https://doi.org/10.1016/S1474-4422(15)00401-9.

13. Lee SK, Lee ST. The laboratory diagnosis of autoimmune encephalitis. J Epilepsy Res. 2016;6(2):45–50. https://doi.org/10.14581/jer.16010.

14. Li L, Sun L, Du R, et al. Application of the 2016 diagnostic approach for autoimmune encephalitis from lancet neurology to Chinese patients. BMC Neurol. 2017;17:195. https://doi.org/10.1186/s12883-017-0974-3.

15. Baumgartner A, Rauer S, Hottenrott T, et al. Admission diagnoses of patients later diagnosed with autoimmune encephalitis. J Neurol. 2019;266(1):124–32. https://doi.org/10.1007/s00415-018-9105-3.

16. Giordano A, Fazio R, Gelibter S, et al. Diagnosing autoimmune encephalitis in a real-world single-centre setting. J Neurol. 2020;267(2):449–60. https://doi.org/10.1007/s00415-019-09607-3.

17. Wagner JN, Kalev O, Sonnberger M, Krehan I, von Oertzen TJ. Evaluation of clinical and paraclinical findings for the differential diagnosis of autoimmune and infectious encephalitis. Front Neurol. 2018;9:434. https://doi.org/10.3389/fneur.2018.00434.

18. Ohkawa T, Fukata Y, Yamasaki M, et al. Autoantibodies to epilepsy-related LGI1 in limbic encephalitis neutralize LGI1-ADAM22 interaction and reduce synaptic AMPA receptors. J Neurosci. 2013;33(46):18161–74. https://doi.org/10.1523/JNEUROSCI.3506-13.2013.

19. Armangue T, Leypoldt F, Malaga I, et al. Herpes simplex virus encephalitis is a trigger of brain autoimmunity. Ann Neurol. 2014;75(2):317–23. https://doi.org/10.1002/ana.24083.

20. Berger B, Pytlik M, Hottenrott T, Stich O. Absent anti-*N*-methyl-D-aspartate receptor NR1a antibodies in herpes simplex virus encephalitis and varicella zoster virus infections. Int J Neurosci. 2017;127(2):109–17. https://doi.org/10.3109/00207454.2016.1147447.

21. Pagad NS, Pradeep N. Clinical named entity recognition methods: an overview. In: Khanna A, Gupta D, Bhattacharyya S, Hassanien AE, Anand S, Jaiswal A, editors. International conference on innovative computing and communications. Singapore: Springer; 2022; pp 151–65.

22. Ceri S, Bozzon A, Brambilla M, Della Valle E, Fraternali P, Quarteroni S. An introduction to information retrieval. In: Ceri S, Bozzon A, Brambilla M, Della Valle E, Fraternali P, Quarteroni S, editors. Web information retrieval. Berlin: Springer; 2013. p. 3–11.

23. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. San Diego: Association for Computational Linguistics; 2016. p. 260–70.

24. Ying Q, Yingfei C. Research of clinical named entity recognition based on Bi-LSTM-CRF. J Shanghai Jiaotong Univ. 2018;23:392–7.

25. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the north American chapter of the association for computational linguistics (NAACL). 2019. p. 4171–4186.

26. Naseem U, Musial K, Eklund P, Prasad M. Biomedical named-entity recognition by hierarchically fusing BioBERT representations and deep contextual-level word-embedding. In: 2020 International joint conference on neural networks (IJCNN). 2020. p. 1–8.

27. Fan B, Fan W, Smith C, Garner HS. Adverse drug event detection and extraction from open data: adeep learning approach. Inform Process Manag. 2020;57(1): 102131. https://doi.org/10.1016/j.ipm.2019.102131.

28. Pintas JT, Fernandes LAF, Garcia ACB. Feature selection methods for text classification: a systematic literature review. Artif Intell Rev. 2021;54(8): 6149–200. https://doi.org/10.1007/s10462-021-09970-6.

29. Deng H, Wang J, Liu X, Liu B, Lei J. Evaluating the outcomes of medical informatics development as a discipline in China: a publication perspective. Comput Methods Programs Biomed. 2018;164: 75–85. https://doi.org/10.1016/j.cmpb.2018.07.001.

30. Spasic I, Uzuner O, Zhou L. Emerging clinical applications of text analytics. Int J Med Inform. 2020;134: 103974. https://doi.org/10.1016/j.ijmedinf.2019.103974.

31. Percha B. Modern clinical text mining: a guide and review. Annu Rev. 2021;4(1):165–87. https://doi.org/10.1146/annurev-biodatasci-030421-030931.

32. Zuliani LA, Zoccarato MB, Gastaldi MC, et al. Diagnostics of autoimmune encephalitis associated with antibodies against neuronal surface antigens. Neurol Sci. 2017;38(Suppl 2):225–9.

33. Zoccarato MA, Gastaldi MB, Zuliani LC, et al. Diagnostics of paraneoplastic neurological syndromes. Neurol Sci. 2017;38(Suppl 2):237–42.

34. Zhang S, Kang T, Zhang X, Wen D, Elhadad N, Lei J. Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models. J Biomed Inform. 2016;60:334–41. https://doi.org/10.1016/j.jbi.2016.02.011.

35. Yang J, Guan Y, He B, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records. J Softw. 2016;27(11): 2725–46.

36. Santiso S, Perez A, Casillas A, Oronoz M. Neural negated entity recognition in Spanish electronic health records. J Biomed Inform. 2020;105: 103419. https://doi.org/10.1016/j.jbi.2020.103419.

37. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781.

38. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. 2013. arXiv:1310.4546. https://doi.org/10.48550/arXiv.1310.4546.

39. Pennington. GloVe: Global vectors for word representation. In: 2014 Conference on empirical methods in natural language processing, EMNLP 2014. Doha, Qatar2014.

40. Lever J, Krzywinski M, Altman N. Points of significance: model selection and overfitting. Nat Methods. 2016;39(9):703–4.

41. Kingma D, Ba J. Adam: a method for stochastic optimization. In: International conference on learning representations. 2014.

42. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5): 552–6. https://doi.org/10.1136/amiajnl-2011-000203.

43. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc. 2014;21(5):808–14. https://doi.org/10.1136/amiajnl-2013-002381.

44. Tang B, Wang X, Yan J, Chen Q. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF. BMC Med Inform Decis Mak. 2019;19(Suppl 3):74. https://doi.org/10.1186/s12911-019-0787-y.

45. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:6.

46. VanderPlas J. Python data science handbook: essential tools for working with data. O'Reilly Media; 2016.

47. Sarah Guido ACM. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media; 2016.

48. Abadi Mn, Barham P, Chen J. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX symposium on operating systems design and implementation (OSDI '2016). 2016.

49. Pang B, Nijkamp E, Wu YN. Deep learning with TensorFlow: a review. J Educ Behav Stat. 2020;45(2): 227–48.

50. Granerod J, Ambrose HE, Davies NW, et al. Causes of encephalitis and differences in their clinical presentations in England: a multicentre, population-based prospective study. Lancet Infect Dis. 2010;10(12):835–44. https://doi.org/10.1016/S1473-3099(10)70222-X.

51. Armangue T, Spatola M, Vlagea A, et al. Frequency, symptoms, risk factors, and outcomes of autoimmune encephalitis after herpes simplex encephalitis: a prospective observational study and retrospective analysis. Lancet Neurol. 2018;17(9): 760–72. https://doi.org/10.1016/S1474-4422(18)30244-8.

52. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. 2019. arXiv preprint arXiv:1910.03771. https://doi.org/10.48550/arXiv.1910.03771.

53. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. 2019. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.

54. Clark K, Luong M-T, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. 2020. arXiv preprint arXiv:2003.10555. https://doi.org/10.48550/arXiv.2003.10555.

55. Cui Y, Che W, Liu T, Qin B, Wang S, Hu G. Revisiting pre-trained models for Chinese natural language processing. Online: association for computational linguistics. 2020. p. 657–68. https://doi.org/10.18653/v1/2020.findings-emnlp.58.

56. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36: 1234–40.

57. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. 2020. arXiv:2006.11371. https://doi.org/10.48550/arXiv.2006.11371.

58. McCoy LG, Brenna CTA, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol. 2022;142:252–7. https://doi.org/10.1016/j.jclinepi.2021.11.001.

59. Cutillo CM, Sharma KR, Foschini L, et al. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med. 2020;3(1):47. https://doi.org/10.1038/s41746-020-0254-2.

60. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise QC. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. https://doi.org/10.1186/s12911-020-01332-6.

61. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

62. Szczepanski M, Pawlicki M, Kozik R, Choras M. New explainability method for BERT-based model in fake news detection. Sci Rep. 2021;11(1):23705. https://doi.org/10.1038/s41598-021-03100-6.

63. Choraś M, Pawlicki M, Puchalski D, Kozik R. Machine learning—the results are not the only thing that matters! What about security, explainability and fairness? In: Krzhizhanovskaya VV, Závodszky G, Lees MH, et al., editors. Computational science—ICCS 2020. Cham: Springer; 2020. p. 615–28.