# Text Classification towards Detecting Misdiagnosis of an Epilepsy Syndrome in a Pediatric Population

**Ryan Sullivan, MS[1], Robert Yao[1], Randa Jarrar, MD[2]**
**Jeffrey Buchhalter, MD[3], PhD, Graciela Gonzalez, PhD[1]**
**[1]Arizona State University, Phoenix, AZ; [2]Phoenix Children's Hospital, Phoenix, AZ;**
**[3]Alberta Children's Hospital, Alberta, Canada**

## Abstract

*When attempting to identify a specific epilepsy syndrome, physicians are often unable to make or agree upon a diagnosis. This is further complicated by the fact that the current classification and diagnosis of epilepsy requires specialized training and the use of resources not typically available to the average clinician, such as training to recognize specific seizure types and electroencephalography (EEG)[1–4]. Even when training and resources are available, expert epileptologists often find it challenging to identify seizure types and to distinguish between specific epilepsy syndromes[5]. Information relevant to the diagnosis is present in narrative form in the medical record across several visits for an individual patient. Our ultimate goal is to create a system that will assist physicians in the diagnosis of epilepsy. This paper explores, as a baseline, text classification methods that attempt to correlate the narrative text features to the diagnosis of West syndrome (Infantile Spasms), using data from Phoenix Children's Hospital (PCH). We tested these methods against a dataset containing known (coded) diagnosis of West Syndrome, and found the best performing method to have a precision / recall / f-measure of 76.8 % / 66.7 % / 71.4 % when evaluated with 10-fold cross validation.*

## Introduction

Epilepsy and misidentification of specific epileptic syndromes has negative implications to public health. Epilepsy affects children and older adults, with 1 in 26 people suffering from it at some point in their lifetime, and about 65 million people affected world-wide[6]. In the U.S., epilepsy is the fourth most common neurologic disorder with a prevalence of 2.2 million and an incidence rate of 150,000 individuals annually. Epilepsy is often poorly managed, misdiagnosed and, in the worst cases, untreated[1,6,7]. These epileptic patients may experience difficulties with independent living that includes difficulties in school, uncertainties about employment possibilities, and limitations on driving. About 1 in 10,000 newly diagnosed patients suffer sudden unexpected death in epilepsy[6]. While seizures, epilepsy, and their sequelae have always been present, there have been recent advances in the knowledge and understanding of the disease, its management, and its treatment. Unfortunately, current diagnostic and treatment methods do not adequately capitalize on these advances and patients who could be helped continue to suffer needlessly.

Epilepsy is a complex neurological disorder that manifests as two or more unprovoked seizures of varying types occurring within a 24-hour period[2,8]. An epileptic seizure, also known as an ictal event, is a transient occurrence of signs and/or symptoms that are the manifestations of abnormally excessive synchronous activity of a set of neurons in the brain[7 9,10]. A specific epilepsy syndrome is characterized by a cluster of signs and symptoms that define a unique epilepsy condition and often includes various seizure types[9–12]. To make the diagnosis of epilepsy, clinicians currently utilize the 1981 and 1989 International League Against Epilepsy (ILAE) classifications and clinical case experience. The identification of a specific epileptic syndrome begins with a description of symptoms by the patient and signs by eyewitnesses and requires the inclusion of electroencephalographic recordings of the ictal events. Depending on the type of seizures a person is suffering, along with other diagnostic measures such as age, developmental history, and EEG data, different types of epilepsy syndromes can be diagnosed, each of which requires different treatments. Unfortunately, few clinicians receive the training of expert epileptologists and epilepsy remains unidentified or misdiagnosed in as many as 12 to 23% of all cases leading to a mismatch of epilepsy syndrome to appropriate treatment and management[13].

Concern over epilepsy misdiagnosis has reached a tipping point in recent years, leading to the Institute of Medicine (IOM) issuing the report, "Epilepsy Across the Spectrum"[6]. In the report, it was recognized that the inability of physicians to make or agree upon a diagnosis consistently for a specific epilepsy syndrome is the major cause of inadequate management of the disease, which in turn, has negatively impacted public health.

As a first step towards better methods for epilepsy syndrome identification, a retrospective analysis was conducted of 27,524 patient records referred to Phoenix Children's Hospital (PCH). The objective of this study is to evaluate text classification methods that do not require any additional annotations, and that can correlate the narrative text present in the records (including EEG reports) to a specific diagnosis (West Syndrome).

**Background**

West Syndrome, also known as infantile spasms was one of the first epilepsy syndromes discovered. It has an incidence rate of 2 to 3.5 per 10,000 live births with 90% of cases occurring in the first year[14]. The diagnosis can be made based on age, developmental history, semiology (observed signs), and EEG patterns. Spasms typically have a neonatal onset and affect boys slightly more than girls between 4 to 8 months of age, but occasionally late onset may occur[15]. Additionally, the child typically has marked developmental delay and mental retardation[14]. Semiologically, a cluster of motor seizures of two major types (spasms and tonic contractions) occur for a duration of less than 1-10 minutes[14,15]. The initial component consists of 2-100 brief epileptic spasms of variable frequency of 1-2 seconds for each spasm[14,15]. These affect primarily the axial muscles of the neck and trunk and appear as characteristic head noddings or "bobbings"[15]. As these seizures occur, various characteristic features on an EEG can be observed. These observations are typically described in detail within the narrative portions of the record and the EEG report.

The clinical features of infantile spasms are considered characteristic, and a diagnosis can be easily made by an expert epileptologist. It is one of the least misdiagnosed epilepsy syndromes, and is thus an excellent case study for the methods proposed, as we expect that most of the true positives will be correctly coded for West Syndrome in the PCH dataset. Still, manual review of false positives is ongoing.

However, it is possible that if certain features co-occur, there may be cases of West Syndrome that may have been missed in the original assessment, and there might be cases that require only one or two additional pieces of information to clinch the diagnosis. An automatic method that could notify the tending physician at the right time could help identify such cases.

Natural Language Processing (NLP) is a subfield of Computer Science and Artificial Intelligence that focuses on human language, and includes tasks such as information extraction and entity recognition. NLP based Clinical Decision Support has been discussed in Demner-Fushman et al., however Meystre et al. suggest that this area of research is relatively underdeveloped compared to other areas of BioNLP[16,17]. However, there do exist a few similar systems. Yetisgen-Yildiz et al. present a system for the identification of patients with acute lung injury from free-text chest x-ray reports[18]. This system uses a feature set based on unigrams, bigrams and trigrams as well as an assertion analysis system to classify the free-text of x-ray reports. Their best performing classifier configuration achieved a precision / recall / f-measure score of 81.7% / 75.6 % / 74.6 %.

Another system developed by Wagholikar et al. uses NLP techniques to generate cervical cancer screening guidelines from free-text Pap reports[19]. This system used hand-crafted rules to suggest cervical cancer screenings based on the text of the Pap reports, and in their evaluation they found that their system suggested the optimal screening recommendations in 73 of their 74 test cases.

The automatic coding of medical text is another related area of research that has been supported by the BioNLP community. This area of study was the subject of a shared task, which challenged teams to automatically assign ICD-9-CM codes to radiology reports[20]. While a number of the top performing systems in the challenge rely on experts to manually craft rules for their systems, Farkas and Szarvas present an approach that uses machine learning techniques to automatically generate coding rules for their system[21]. Their resulting system achieves a precision / recall / f-measure score of 87.6% / 90.0 % / 88.9 % on the shared task test set, which is competitive with the highest scoring system of the challenge, which achieved an f-measure of 89.1 %.

**Methods**

*Dataset:* A retrospective analysis was conducted of 27,524 patient records referred to Phoenix Children's Hospital (PCH). These records consist of patients that have been coded for epilepsy (all ICD9 345 codes) as well as those with insufficient clinical evidence to support a specific diagnosis of an epilepsy syndrome (patients coded with ICD9 780.39 'Other Convulsions'). We divided the patient records into three groups: 1) 144 patients coded for Infantile Spasms (ICD9 codes 345.60 and 345.61); 2) 2,818 patients with records that contain Infantile Spasm-related keywords ["infantile spasms", "tuberous sclerosis", "hypsarrythmia", "ACTH", "prednisolone" and "aicardi syndrome"], but are not coded for Infantile Spasms; and 3) 24,562 patients with records neither coded for infantile spasms nor containing Infantile Spasm-related keywords. For our experiments, we created a corpus consisting of the records of the 144 patients coded for infantile spasms as positive examples and the records of 3,600 randomly chosen patients from group three as negative examples.

*Preprocessing:* For this classification task, we only used the free-text from discharge summaries and EEG reports. We used a simple tokenizer to tokenized the text at whitespace and punctuation, and removed all digits and special characters. We also removed all Infantile Spasm-related keywords and English stopwords from the text.

*Feature generation:* We tested two different techniques for generating a feature set from the patient record free-text, TF-IDF vectors and a topic distribution based on Latent Dirichlet Allocation (LDA).

Our first feature set consisted of TF-IDF (term frequency-inverse document frequency) vectors, a popular scheme that is generally used for indexing documents for information retrieval[22]. For each term in a document, the term frequency (how many time a term appears in the document) and the inverse document frequency (a measure of how rare a term is across documents) is calculated, and these values are used to construct a term vector representation of the document.

Our second approach consisted of representing each patient as a topic distribution based on Latent Dirichlet Allocation (LDA)[23]. LDA is an unsupervised technique used for topic discovery and text classification. It assumes that each document is generated based on some topic distribution and each topic's word distribution. It then attempts to use the observed information, i.e. the words in the documents, to predict the unobserved information, i.e. the topics, the topic distribution and the word distributions within each topic. We trained an LDA topic model on the patient data consisting of 1,500 topics, with the topic size chosen based on testing a subsample of the corpus, and we used the corpus to estimate the topic distribution for each record. We used the Mallet toolkit to build the LDA models[24] We used this topic distribution as a feature set; in a method tantamount to using LDA for dimensionality reduction, and a method mentioned in Blei and McAuliffe[25].

*Training data sampling:* One issue we encountered with the dataset is the relative rarity of Infantile Spasm patients compared to the negative examples. Because of this disparity, we ran into issues of having too few positive examples to train our classifiers.

We tested two solutions to this problem, oversampling the positive classes and undersampling the negative classes. In both cases, we attempted to have a 3 to 1 ratio of negative to positive examples (compared to the 25 to 1 ratio of our corpus). In oversampling, we keep the number of negative examples the same, but give the positive examples a weight of 8.33 times the negative examples. In undersampling, we trained the model on a random subset of the data which has the 3 to 1 ratio.

*Classification and Evaluation* We compared two different classification algorithms for this task; a multinomial Naïve Bayes classifier and a Support Vector Machine classifier[26,27]. These classifiers were chosen for their speed and for their documented performance in text classification tasks.

We evaluated our classification models using 10-fold cross validation, and we trained and evaluated each classification model using Weka, and as specified as the Weka documentation, the folds were selected randomly[28].

**Results and Discussion**

The results of our experiment can be seen in Table 1.

These results do not overtake the top performing systems, yet they are compatible to other systems in this domain, and they represent a reasonable baseline for which to continue research. Furthermore, these results show that the use of domain knowledge is not a necessary requirement to achieve reasonable results. There are also a few conclusions that

Table 1: Classification results.

| **Classifier** | *Precision* | *Recall* | *F-Measure* |
|---|---|---|---|
| multinomial Naïve Bayes - LDA - no sampling | 16.5% | 61.8% | 26.1% |
| multinomial Naïve Bayes - TF*IDF - no sampling | 19.4% | 53.5% | 28.5% |
| SVM - LDA - no sampling | 45.0% | 34.7% | 39.2% |
| SVM - TF*IDF - no sampling | 55.1% | 29.9% | 38.7% |
| multinomial Naïve Bayes - LDA - oversampling | 47.4% | 75.7% | 58.3% |
| multinomial Naïve Bayes - TF*IDF - oversampling | 77.5% | 55.6% | 64.7% |
| SVM - LDA - oversampling | 81.4% | 29.9% | 43.7% |
| SVM - TF*IDF - oversampling | 88.2% | 29.9% | 44.6% |
| multinomial Naïve Bayes - LDA - undersampling | 50.0% | 70.8% | 58.6% |
| multinomial Naïve Bayes - TF*IDF - undersampling | 59.7% | 75.0% | 66.5% |
| SVM - LDA - undersampling | 70.1% | 65.3% | 67.6% |
| SVM - TF*IDF - undersampling | 76.8% | 66.7% | 71.4% |

can be drawn from these results. The first conclusion is that a sampling method (either over or under sampling) is a requirement to get usable performance. However, from analyzing the results, it seems that oversampling may be over fitting the data, which causes the precision-recall difference between the over and under sampling.

We can also gather from these results that LDA is an effective dimension reduction technique for this type of text, when using a SVM classifier. Though the TF-IDF features outperform the LDA-based features, the TF-IDF vectors are 10 times as large, and represent a non-trivial computation time cost for certain classifiers. Finally, we can see that SVM consistently had high precision-low recall compared to the Naïve Bayes' comparatively low precision-high recall. Though one would ideally want both high precision and high recall, these classifier tendencies are worth keeping in mind depending on the task.

**Conclusion**

The use of Natural Language Processing for Clinical Decision Support is an academically underserved domain that holds a large potential. The results of our baseline system are comparable to other systems in the domain, and can identify misdiagnosed epilepsies or improve the ability for medical doctors to identify an epilepsy syndrome not previously diagnosed. This can potentially improve the match between patient and the best treatment available for a specific epilepsy syndrome, and can further the goal for better seizure control and improvement in quality of life for the patient.

# References


1. Birbeck, G.L.. Revising and refining the epilepsy classification system: Priorities from a developing world perspective. Epilepsia 2012;53(s2):18–21.
2. Berg, A.T., Berkovic, S.F., Brodie, M.J., Buchhalter, J., Cross, J.H., Van Emde Boas, W., et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ilae commission on classification and terminology, 2005–2009. Epilepsia 2010;51(4):676–685.
3. Scheffer, I.E.. Epilepsy: A classification for all seasons? Epilepsia 2012;53(s2):6–9.
4. Berg, A.T., Cross, J.H.. Towards a modern classification of the epilepsies? The Lancet Neurology 2010;9(5):459–461.
5. Ottman, R., Hauser, W.A., Stallone, L.. Semistructured interview for seizure classification: agreement with physicians' diagnoses. Epilepsia 1990;31(1):110–115.

6. England, M.J., Liverman, C.T., Schultz, A.M., Strawbridge, L.M.. Epilepsy across the spectrum: Promoting health and understanding.: A summary of the institute of medicine report. Epilepsy & Behavior 2012;25(2):266–276.

7. Thurman, D.J., Beghi, E., Begley, C.E., Berg, A.T., Buchhalter, J.R., Ding, D., et al. Standards for epidemiologic studies and surveillance of epilepsy. Epilepsia 2011;52(s7):2–26.

8. Blume, W.T., Lüders, H.O., Mizrahi, E., Tassinari, C., van Emde Boas, W., Engel, J.. Glossary of descriptive terminology for ictal semiology: report of the ilae task force on classification and terminology. Epilepsia 2001;42(9):1212–1218.

9. Engel, J.. A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ilae task force on classification and terminology. Epilepsia 2001;42(6):796–803.

10. Engel Jr, J.. Ilae classification of epilepsy syndromes. Epilepsy research 2006;70:5–10.

11. Epilepsy, A.. Proposal for revised classification of epilepsies and epileptic syndromes. The treatment of epilepsy: principles & practice 2006;354.

12. Wolf, P.. Basic principles of the ilae syndrome classification. Epilepsy research 2006;70:20–26.

13. Network, S.I.G.. Diagnosis and management of epilepsies in children and young people; guideline no. 81 ed. Royal College of Physicians; 2005.

14. Pellock, J.M., Hrachovy, R., Shinnar, S., Baram, T.Z., Bettis, D., Dlugos, D.J., et al. Infantile spasms: a us consensus report. Epilepsia 2010;51(10):2175–2189.

15. Shields, W.D.. Infantile spasms: little seizures, big consequences. Epilepsy Currents 2006;6(3):63–69.

16. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.. What can natural language processing do for clinical decision support? Journal of biomedical informatics 2009;42(5):760–772.

17. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;35:128–44.

18. Yetisgen-Yildiz, M., Bejan, C.A., Wurfel, M.M.. Identification of patients with acute lung injury from free-text chest x-ray reports ????;.

19. Wagholikar, K.B., MacLaughlin, K.L., Henry, M.R., Greenes, R.A., Hankey, R.A., Liu, H., et al. Clinical decision support with automated text processing for cervical cancer screening. Journal of the American Medical Informatics Association 2012;19(5):833–839.

20. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K.B., et al. A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Association for Computational Linguistics; 2007, p. 97–104.

21. Farkas, R., Szarvas, G.. Automatic construction of rule-based icd-9-cm coding systems. BMC bioinformatics 2008;9(Suppl 3):S10.

22. Salton, G., McGill, M.J.. Introduction to moderm information retrieval. 1983.

23. Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent dirichlet allocation. the Journal of machine Learning research 2003;3:993–1022.

24. McCallum, A.K.. Mallet: A machine learning for language toolkit 2002;.

25. Blei, D.M., McAuliffe, J.D.. Supervised topic models. arXiv preprint arXiv:10030783 2010;.

26. McCallum, A., Nigam, K., et al. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization; vol. 752. Citeseer; 1998, p. 41–48.

27. Platt, J.C.. 12 fast training of support vector machines using sequential minimal optimization 1999;.

28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.. The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 2009;11(1):10–18.