Language Model Applications for Early Diagnosis of Childhood Epilepsy

Jitse Loyens[1], Geertruida Slinger[2], Nynke Doornebal[3], Kees P.J. Braun[2], Willem M. Otte[2]#, Eric van Diessen[2,4]#

Affiliations

1. Faculty of Medicine, Utrecht University, the Netherlands
2. Department of Child Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht and Utrecht University, The Netherlands
3. Department of Pediatrics, Martini Hospital, Groningen, the Netherlands
4. Department of Pediatrics, Franciscus Hospital, Rotterdam, the Netherlands

# Authors contributed equally

Corresponding author: Eric van Diessen Department of Child Neurology, UMC Utrecht Brain Center University Medical Center Utrecht and Utrecht University Room KG 01.310.0, P.O. Box 85090 Utrecht 3508 AB, The Netherlands Email: E.vanDiessen-3@umcutrecht.nl

Number of words abstract: 291

Number of words Main text: 3144

Number of Figures / Tables: 3 / 1

ABSTRACT

Objective: Accurate and timely epilepsy diagnosis is crucial to reduce delayed or unnecessary treatment. While language serves as an indispensable source of information for diagnosing epilepsy, its computational analysis remains relatively unexplored. This study assessed – and compared – the diagnostic value of different language model applications in extracting information and identifying overlooked language patterns from first-visit documentation to improve the early diagnosis of childhood epilepsy.

Methods: We analyzed 1,561 patient letters from two independent first seizure clinics. The dataset was divided into training and test sets to evaluate performance and generalizability. We employed two approaches: an established Naïve Bayes model as a natural language processing technique, and a sentence-embedding model based on the Bidirectional Encoder Representations from Transformers (BERT)-architecture. Both models analyzed anamnesis data only. Within the training sets we identified predictive features, consisting of keywords indicative of 'epilepsy' or 'no epilepsy'. Model outputs were compared to the clinician's final diagnosis (gold standard) after follow-up. We computed accuracy, sensitivity, and specificity for both models.

Results: The Naïve Bayes model achieved an accuracy of 0.73 (95% CI: 0.68-0.78), with a sensitivity of 0.79 (95% CI: 0.74-0.85) and a specificity of 0.62 (95% CI: 0.52-0.72). The sentence-embedding model demonstrated comparable performance with an accuracy of 0.74 (95% CI: 0.68-0.79), sensitivity of 0.74 (95% CI: 0.68-0.80), and specificity of 0.73 (95% CI: 0.61-0.84).

Conclusion: Both models demonstrated relatively good performance in diagnosing childhood epilepsy solely based on first-visit patient anamnesis text. Notably, the more advanced sentence-embedding model showed no significant improvement over the computationally simpler Naïve Bayes model. This suggests that modeling of anamnesis data does depend on

word order for this particular classification task. Further refinement and exploration of language models and computational linguistic approaches are necessary to enhance diagnostic accuracy in clinical practice.

1. INTRODUCTION

Epilepsy significantly impacts psychosocial well-being and can adversely affect health-related quality of life.[1,2] This impact is particularly concerning in children, where recurrent seizures can interfere with normal brain development, potentially leading to cognitive and behavioral impairments.[3–5] These serious consequences underscore the critical importance of obtaining an early and accurate diagnosis of epilepsy.

Diagnosing epilepsy presents significant challenges due to its polymorphic nature.[6–8] Research has shown that nearly half of the patients assessed for initial seizures were already experiencing recurrent, undiagnosed seizures at the time of evaluation. While diagnostic time is typically brief for clearly identifiable cases of epilepsy, it can extend beyond a year for complex or ambiguous presentations.[9,10] This diagnostic uncertainty can have serious consequences: diagnostic delays expose children to ongoing seizures that may impair cognitive development, while false-positive diagnoses can lead to unnecessary administration of antiseizure medications with potential adverse effects.[11–13]

Language plays a fundamental role in epilepsy diagnosis, treatment evaluation, and patient care management. Clinicians rely heavily on patient history and narrative to distill relevant clinical information.[14] This makes collected text a rich and versatile medium for gaining deep insight into the patient's condition—an essential component for a comprehensive approach to epilepsy care. Despite advances in ancillary investigations, clinical information from patient records remains indispensable for diagnosing and monitoring epilepsy.[15–17] However, this wealth of information is often stored in electronic health records in an unstructured manner, limiting its optimal utilization in clinical decision-making.[18]

The emergence of natural language processing (NLP) offers a promising solution for systematically processing this unstructured textual data. NLP, a form of artificial intelligence,

specializes in the computational analysis of spoken and written language to identify general patterns and trends and extract relevant information.[19–21] This involves converting unstructured text into a structured format and applying computational algorithms to analyze these structured features, enabling the retrieval of desired information.

In epilepsy research, there is a growing trend toward NLP applications, including patient identification, risk stratification, and outcome prediction. In clinical settings, NLP can contribute significantly to the early detection and classification of medical conditions, thereby reducing time to diagnosis and treatment.[18] Recent advances have led to improved NLP models with new generative properties, known as large language models (LLMs).[22–24] The essence of these models is a transformer architecture with an attention-layer, allowing both an efficient representation and retrieval of relevant information in (textual) data.[24] Despite the potential of these more advanced language models, their applicability for early diagnosis of epilepsy based on medical documentation remains limitedly explored.[25–27] This study aims to assess—and compare—the diagnostic value of different NLP approaches using medical letters from first consultations to facilitate the early diagnosis of childhood epilepsy.

2. METHODS

2.1. Dataset

Our analysis encompassed 1,561 medical patient letters, with 1,250 originating from University Medical Center Utrecht (UMCU) and 311 from Martini Hospital Groningen (MZG). We retrospectively collected data from children (age < 18 years) referred to the First Seizure Clinic (FSC) between 2008 and May 2022. These data were originally collected for previously published studies focusing on prediction model development for childhood epilepsy and the clinical characteristics and diagnoses of children referred to an FSC.[28,29] The

institutional ethics committee of both University Medical Center Utrecht and Martini Hospital approved the use of anonymized retrospective data for research purposes with- out informed consent.

All patient letters were either written and/or supervised by experienced pediatric neurologists. For each patient, we included both the initial diagnosis (established after FSC consultation) and the final diagnosis (reached through consensus among doctors and/or ancillary investigations at the latest follow-up, recorded within a two-year period). Follow-up occurred for children with inconclusive diagnoses at the first consultation and for those initially diagnosed with epilepsy. Children whose epilepsy diagnosis was ruled out were referred back to their referral specialist or general practitioner for follow-up.

Both initial and final diagnoses were categorized into three groups: ‘epilepsy’, ‘no epilepsy’, and ‘unclear’ (**Figure 1**) and served as the model’s outcome. All epilepsy diagnoses were established according to the International League Against Epilepsy definition of epilepsy.[30] A diagnosis was classified as ‘unclear’ at the initial stage if ancillary investigations were deemed necessary to confirm or reject the epilepsy diagnosis. The final diagnosis was classified as ‘unclear’ if, despite further investigations, uncertainty remained about whether the events were indeed epilepsy-related.[7]

2.2. Study Design

We conducted a retrospective analysis of the letters to assess the clinical value of language models for early diagnosis of childhood epilepsy. This was achieved through binary text classification, specifically by training classification models based on textual features and predicting the class of new texts within the ‘epilepsy’ and ‘no epilepsy’ patients. To reduce interpretative bias, we exclusively used textual information from patient anamnesis,

excluding subjective information from ancillary investigations, conclusions, treatment plans, and clinical considerations.

Our study involved two distinct analyses. **Analysis A**: we combined data from both hospitals and randomly divided it into a training set (80%; 1,173 subjects) and a test set (20%; 293 subjects). To ensure representative distribution of final diagnoses in both sets, we applied stratification based on the final diagnosis groups. **Analysis B**: we created a separate test set comprising all 316 subjects that remained unclear after initial FSC evaluation. This second analysis aimed to determine whether the model could accurately classify initially unclear cases as either having epilepsy or not (**Figure 1**).

The letter corpus exhibited considerable variation in textual length, ranging from 63 to 1,070 words, with a median of 400 words and a mean of 414 words. Four cases (three from the UMC Utrecht; two male subjects) were excluded from the training set due to their succinct nature, consisting of only single sentences in their amnestic report.

2.3 Naïve Bayes Model

We used a Naïve Bayes classifier as NLP approach, giving its simplicity and effectiveness in text classification. The essence of the model is the application of Bayes' theorem, assuming a strong independence between features.[31,32] Model development contained three phases: data preprocessing, data analysis with feature selection, and classification (**Figure 2**).

Data preprocessing – Preprocessing encompasses several key steps including corpus creation, tokenization, data cleaning, lowercasing, n-gram generation, and stop word removal. Creating a corpus involves collecting and organizing a substantial amount of textual data in a structured manner to facilitate systematic analysis and processing. Text was then divided into tokens (i.e., words) through tokenization. Undesired characters, such as punctuation marks, symbols, URLs, and separators, were omitted (data cleaning). Lowercasing converted all

characters in the text to lowercase letters, ensuring consistency across the tokens. Afterwards, n-grams were generated, with a maximum n-value of 2. N-grams are sequences of consecutive words and will be used as features for the text classification model. It was decided to generate unigrams (single words such as "trekkingen" *("jerks")*) and bigrams (pairs of consecutive words such as "geen_trekkingen" *("no_jerks")*). The final step involved removing stop words from the generated n-grams. Removing stop words after generating n-grams ensures that some meaningful bigrams are retained, even if they contain stop words (e.g., "geen_koorts" *("no_fever")* may be retained while "geen" *("no")* and "koorts" *("fever")* may individually be stop words). Stop words contain common words including prepositions, personal pronouns, units, and auxiliary verbs that lack informativeness and may interfere with model development. After the preprocessing step, the dataset was split into training and test sets.

Data analysis – We created a document-feature matrix (DFM) for the training set to enable structured analysis of text data, representing documents (letters) as rows and features (i.e., all n-grams) as columns. The matrix values represented the frequency of a features in each letter, creating a Bag-of-Words (BoW) model.[33] In this model the input text is represented as a collection of words, disregarding the order in which they appear. We applied Term Frequency-Inverse Document Frequency (TF-IDF) to weigh features based on their frequency in individual letters. TF-IDF reduces the influence of frequently occurring features while emphasizing more informative ones.

Feature selection – Feature selection was achieved through Recursive Feature Elimination (RFE) with 5-fold cross-validation. RFE identified the top 300 features that were most informative for the model's performance. A selection of 300 features was based on theoretical and practical reasons. Firstly, we wanted to follow the rule of thumb that recommends one feature per ten cases to minimize overfitting and optimize model

performance. As the dataset is of medium size, we adjusted this rule to one feature per five cases, resulting in the selection of 300 features. Secondly, the literature supports this selection, as studies frequently use between 200 and 300 features to capture significant patterns while minimizing noise, thereby enhancing the robustness and generalizability of the model. Thirdly, fewer features improve computational efficiency, making the model more practical for implementation. Moreover, fewer features improve the model's interpretability and transparency, facilitating a better understanding of which variables contribute to its predictions. As a hyperparameter for the Naive Bayes model, the smoothing parameter (α) was added to prevent zero probabilities.

2.4. Sentence-embedding Model and Subsequent Classification

This study also employed a classification model to predict epilepsy diagnoses based on patient text records that takes—in contrast to the BoW approach—word sequence into account. First, textual data underwent systematic preprocessing. Initial preprocessing steps included case normalization to lowercase, standardization of special characters to their lexical equivalents, removal of extraneous punctuation marks, and normalization of whitespaces. Next, the processed texts were then embedded using a freely-available multilingual embedding (i.e., the paraphrase-multilingual-mpnet-base-v2 transformer model).[*] The embedding model implements the Sentence-BERT architecture to generate a contextualized 768-dimensional semantic vector representations for each text, irrespective of its length.[34] This embedding model was selected for its capacity to preserve both sequential word order information and cross-lingual semantic relationships. Third, the resulting high-dimensional embeddings served as input features for a gradient boosting classifier implemented through the XGBoost framework in R.[35] The binary classification model employed a linear booster

[*] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

with default hyperparameters, leveraging sequential tree building to iteratively optimize the prediction objective while maintaining computational efficiency.

2.5. Performance Evaluation

Performance evaluation utilized confusion matrices to compare actual and predicted classifications through decision statistics from contingency tables. We compared each model's output to the clinician's final diagnosis (gold standard). Key performance metrics included: accuracy (i.e., the proportion of correct classifications), sensitivity (true positive rate), and specificity (true negative rate). All evaluation analyses were performed using R software, version 4.4.0.

3. RESULTS

3.1 Data Characteristics

The median age at the first seizure was 4.5 years (95% CI: 4.0-4.9). The maximum age recorded was 17.8 years, while the minimum age was 1 month. The majority of patients were male, comprising 853 individuals (54.6%). After the first consultation, 366 diagnoses were classified as 'epilepsy', 795 as 'no epilepsy', and 400 as 'unclear'. According to the final diagnoses, 514 diagnoses were classified as 'epilepsy' (413 from UMCU and 101 from MZG), 958 as 'no epilepsy' (767 from UMCU and 191 from MZG), and 89 as 'unclear' (70 from UMCU and 19 from MZG). The data characteristics are presented (**Table 1A, Supplementary Materials**).

3.2. Most Important Features

The *Term Frequency-Inverse Document Frequency* identified several key features most characteristic for the epilepsy classification texts. Notable predictive n-gram feat1ures included: "spray" *("spray")*, "kwijlde" *("drooled")*, "haar_mond" *("her_mouth")*, "insult_doorgemaakt" *("experienced_insult")*, "dubbele tong)" *("slurred_speech")* and "afgelopen_dagen" *("last_days")*. Some features directly reflected clinical observations or descriptions that frequently appear in letters of epilepsy patients, while others, such as "afgelopen_dagen" ("*last_days*") showed less obvious connection to epilepsy. Lists of the most important features, for the epilepsy as well as the control group, are provided in **Figure 3**.

3.3. Classification Model Performance

Analysis A - The performance of the language models was evaluated on a test set of 293 letters. The Naïve Bayes model correctly identified 62 letters as positive and 153 as negative, resulting in 40 false positives and 30 false negatives, with an overall accuracy of 0.73 (95% CI: 0.68-0.78). The Sentence-embedding model correctly identified 40 letters as positive and 176 as negative, classifying 62 false positives and 15 false negatives, with an overall accuracy of 0.74 (95% CI: 0.68-0.79). An overview including sensitivity, specificity, PPV and NPV is provided (**Table 1**).

Analysis B - The performance of the language models was evaluated on a test set of 319 letters with an 'unclear' diagnosis. The Naïve Bayes model correctly identified 60 letters as positive and 173 as negative, resulting in 37 false positives and 46 false negatives, with an overall accuracy of 0.74 (95% CI: 0.69–0.79). The Sentence-embedding model correctly identified 31 letters as positive and 196 as negative, classifying 66 false positives and 23 false negatives, with an overall accuracy of 0.72 (95% CI: 0.67–0.77). An overview including sensitivity, specificity, PPV and NPV is provided (**Table 1**).

4. DISCUSSION

This study evaluated and compared different language model applications for improving early diagnosis of childhood epilepsy through automated analysis of first-visit documentation. Our findings revealed comparable performance between the simpler Naïve Bayes model and the more advanced Sentence-embedding model, with both achieving moderate to good diagnostic accuracy. Notably, both models demonstrated higher specificity than sensitivity across all analyses, suggesting particular utility in helping clinicians rule out epilepsy diagnoses and identify cases requiring additional investigation. Previous research has established the value of NLP in various aspects of epilepsy care, including patient identification,[36–38] information retrieval,[39–41] and coping strategies.[42,43] Recent studies have begun exploring language applications in early clinical phenotyping and genetic epilepsies.[44,45] However, our study uniquely addresses the specific challenges of early childhood epilepsy diagnosis, where textual analysis holds particular promise given the heterogeneous presentation of symptoms.

The performance metrics should be considered in context of the model performance: our models relied solely on patient narratives, deliberately excluding information from EEG reports, clinical evaluations, and medical conclusions. From this perspective, the application of language models could even be used in the early phase of clinical evaluation of child suspected of epilepsy. Interestingly, the transformer-based Sentence-embedding model – which takes word order into account – demonstrates no significant improvement over the Naïve Bayes model. The Naïve Bayes model is regarded as a robust classification model, even when working with limited data and feature sets.[31] Transformer-based language models perhaps require longer text sequences to effectively recognize desired patterns, particularly in

cases with less variation in language utilization. With limited text input, simpler models can offer greater practical value in practice, where speed, simplicity and apprehensibility often take precedence in the implementation within clinical workflows. The achieved accuracy levels suggest potential value for early-stage screening and decision support.

Interestingly, our study revealed that both obvious as less obvious words (or combinations) are of additional value for correct classification. The use of epilepsy-related terminology could reflect the physician's (implicit) evaluation of the clinical case during consultation. Unrelated word (combinations) with no obvious relation to epilepsy that were classified as relevant features for model develop may represent either underlying linguistic patterns common in epilepsy-related letters, or potential limitations in the model's feature selection process. Previous efforts in the field have revealed similar insights into the non-semantic evaluation of patient history, and showed that hesitations and formulation efforts might be of additional value when diagnosing epilepsy.[46,47] Future research efforts should therefor a comparison of different language model approaches to further elucidate the true value of these implicit language information for diagnosing epilepsy.

This study benefits from a substantial and diverse dataset collected from two hospitals, enhancing the robustness and generalizability of the model's results. The retrospective nature and moderate size of our dataset are, however, potential contributors to the limited sensitivity and PPV of both models, thereby increasing the chance of missed epilepsy cases and false positives. Furthermore, performance was significantly higher on the training sets (not reported) compared to the test sets, indicating potential overfitting. Overfitting occurs when the model learns the textual details and noise in the training data, which impairs its generalizability to new data. This can result from excessive noise, an excessive number of features, irrelevant features, or insufficient training data. Equally important to consider is the imbalanced dataset we used (uneven class distribution),

predominantly consisting of letters from children with a 'no epilepsy' diagnosis. Imbalanced data can hinder a model's ability to learn the minority class, as (language) models often exhibit a preference for the majority class.[48]

From a model perspective, few limitations should be mentioned. A Naïve Bayes model is a relatively limited in its capacity to learn complex (textual) relations.[31] The model does not adequately account for word order or combinations of words, potentially resulting in misinterpretations of negations and the overall meaning within clinical text. Confounding factors like typographical errors, abbreviations, double negations, and letters written by multiple authors can adversely affect the classification process. Additionally, RFE was applied to a subset of the top 8000 features (i.e., 300) due to computational constraints, possibly excluding relevant features. A general limitation of feature selection is the possible omission of rare but significant features, particularly in the context of rare diseases or syndromes. Technically, more word-order-oriented models could (partially) overcome the aforementioned model limitations due to their transformer-architecture in which meaningful textual relations are represented internally. Mechanism that drives these models to achieve such model properties remain difficult to grasp, prohibiting a better of understanding of these models.[49,50]

Future research should incorporate a prospective design to explore the clinical applicability. Prospective studies enhance variable control, minimize data noise, and allow real-time language capture, thereby reducing biases and missing data. This would also allow to capture a recorded – instead of a written – patient history that would inevitably lead to new potential hidden language domain sources (e.g., phonology, prosody, syntax use) to improve epilepsy diagnosis.[46,47] This could be particularly beneficial for LLMs as these models excel in retrieving ‘hidden’ textual association that might be use for classification. Enhancing algorithms, refining feature selection, and utilizing larger, more diverse datasets are essential

to improve diagnostic accuracy. Apart from methodological improvements, integration of language-based classification models with existing clinical diagnostic tools in epilepsy care would be a next step to explore its actual clinical value.[28,51,52]

5. CONCLUSIONS

Our study demonstrates that both simple and complex language models can achieve meaningful performance in supporting early childhood epilepsy diagnosis, even when limited to first-visit documentation. The comparable performance between Naïve Bayes and more sophisticated transformer-based language model suggests that simpler, more interpretable models may be preferable for initial clinical applications as long as the input data is limited in size and complexity. While further refinement is needed, these findings support the potential value of computational linguistic approaches in improving early epilepsy diagnosis and patient care. The higher sensitivity demonstrated by both models suggests particular utility in helping clinicians identify cases that do not require additional investigation, potentially streamlining the diagnostic process and reducing unnecessary testing. As these methods continue to evolve, their integration into clinical practice could provide valuable decision support for clinicians while maintaining the essential role of clinical expertise in final diagnostic decisions.

FIGURES AND TABLES

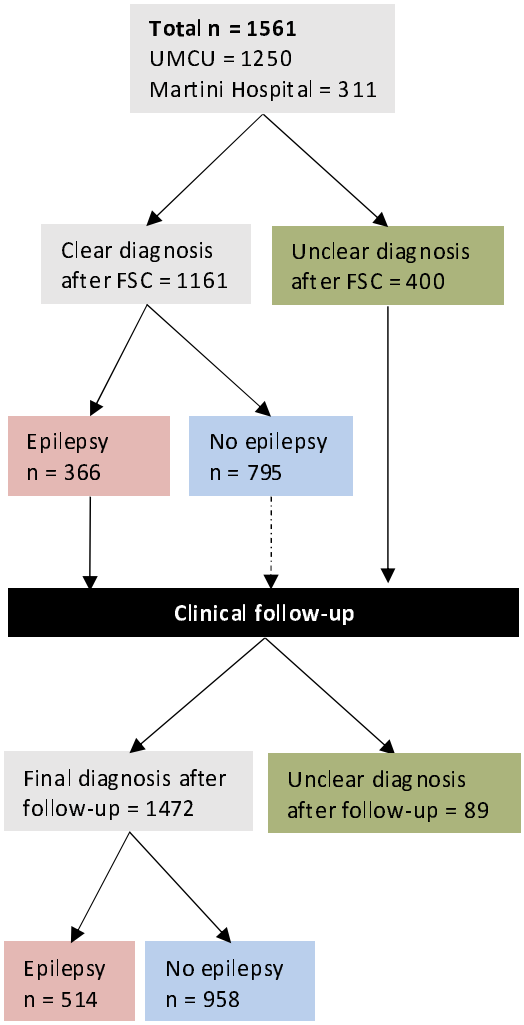

Figure 1. *Flowchart illustrating the diagnostic pathway for children referred to the FSC. The flowchart outlines the process from the first FSC consultation to the final diagnosis, including follow-up procedures. The diagnoses are categorized as 'epilepsy', 'no epilepsy', or 'unclear'.*

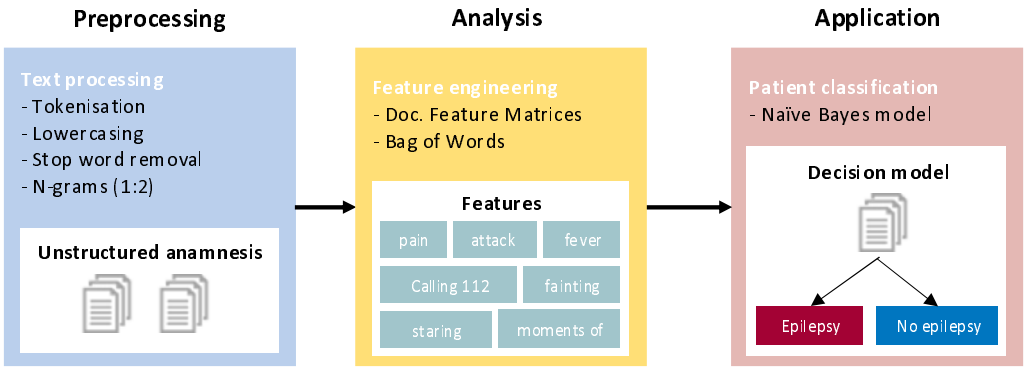

Figure 2. *NLP workflow for classifying 'epilepsy' or 'no epilepsy' diagnosis based on unstructured letters from the first consultations. The process consists of three main stages: preprocessing, analysis, and application.*

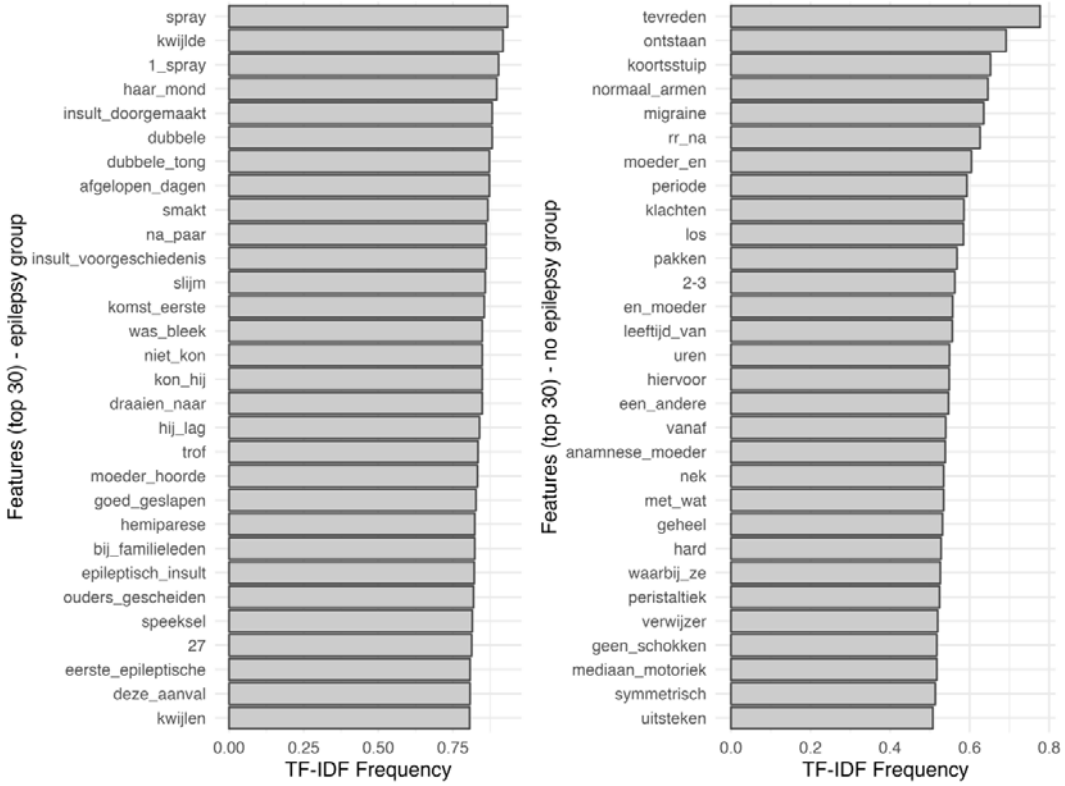

Figure 3. *A graphical representation of the most prevalent features for each condition. TF-IDF = Term Frequency-Inverse Document Frequency; to weigh features based on their frequency in individual letters. A complete translation list in English of features is provided* (**Table 2A, Supplementary Materials**).

Analysis	Model	Accuracy	Sensitivity	Specificity	PPV	NPV
A	NB	0.73 (0.68–0.78)	0.79 (0.74–0.85)	0.62 (0.52–0.72)	0.80 (0.74–0.86)	0.61 (0.51–0.70)
	Sentence-embedding	0.74 (0.68–0.79)	0.74 (0.68–0.80)	0.73 (0.61–0.84)	0.92 (0.88–0.96)	0.39 (0.30–0.49)
B	NB	0.74 (0.69–0.79)	0.82 (0.77–0.88)	0.57 (0.47–0.66)	0.79 (0.74–0.84)	0.62 (0.52–0.72)
	Sentence-embedding	0.72 (0.67–0.77)	0.75 (0.70–0.80)	0.57 (0.44–0.71)	0.89 (0.85–0.94)	0.32 (0.23–0.41)

Table 1. *Model performance metrics for epilepsy diagnosis: test sets. NB = Naïve Bayes.*

Between parentheses = 95% Confidence Interval.

REFERENCES

1. Hamiwka L, Singh N, Niosi J, Wirrell E. Perceived health in children presenting with a "first seizure." Epilepsy Behav. 2008/07/12. 2008; 13(3):485–8.
2. Modi AC, King AS, Monahan SR, Koumoutsos JE, Morita DA, Glauser TA. Even a single seizure negatively impacts pediatric health-related quality of life. Epilepsia. 2009/06/06. 2009; 50(9):2110–6.
3. Camfield PR, Camfield CS, Dooley JM, Tibbles JA, Fung T, Garner B. Epilepsy after a first unprovoked seizure in childhood. Neurology. 1985/11/01. 1985; 35(11):1657–60.
4. Berg AT, Loddenkemper T, Baca CB. Diagnostic delays in children with early onset epilepsy: impact, reasons, and opportunities to improve care. Epilepsia. 2013/12/10. 2014; 55(1):123–32.
5. Holst AG, Winkel BG, Risgaard B, Nielsen JB, Rasmussen PV, Haunsø S, et al. Epilepsy and risk of death and sudden unexpected death in the young: A nationwide study. Epilepsia [Internet]. 2013 [cited 2024]; 54(9):1613–20. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/epi.12328
6. Parviainen L, Kälviäinen R, Jutila L. Impact of diagnostic delay on seizure outcome in newly diagnosed focal epilepsy. Epilepsia Open [Internet]. 2020 [cited 2024]; 5(4):605–10. Available from: https://onlinelibrary-wiley-com.utrechtuniversity.idm.oclc.org/doi/full/10.1002/epi4.12443
7. Slinger G, Noorlag L, van Diessen E, Otte WM, Zijlmans M, Jansen FE, et al. Clinical characteristics and diagnoses of 1213 children referred to a first seizure clinic. Epilepsia Open [Internet]. 2024 [cited 2024]; 9(2):548–57. Available from: https://onlinelibrary-wiley-com.utrechtuniversity.idm.oclc.org/doi/full/10.1002/epi4.12883
8. Pellinen J, French J, Knupp KG. Diagnostic Delay in Epilepsy: the Scope of the Problem. Curr Neurol Neurosci Rep [Internet]. 2021 [cited 2024]; 21(12):1–6. Available from: https://link-springer-com.utrechtuniversity.idm.oclc.org/article/10.1007/s11910-021-01161-8
9. Firkin AL, Marco DJT, Saya S, Newton MR, O'Brien TJ, Berkovic SF, et al. Mind the gap: Multiple events and lengthy delays before presentation with a "first seizure." Epilepsia [Internet]. 2015 [cited 2024]; 56(10):1534–41. Available from: https://onlinelibrary-wiley-com.utrechtuniversity.idm.oclc.org/doi/full/10.1111/epi.13127
10. Kalilani L, Faught E, Kim H, Burudpakdee C, Seetasith A, Laranjo S, et al. Assessment and effect of a gap between new-onset epilepsy diagnosis and treatment in the US. Neurology. 2019; 92(19):E2197–208.
11. Leach JP, Lauder R, Nicolson A, Smith DF. Epilepsy in the UK: Misdiagnosis, mistreatment, and undertreatment?: The Wrexham area epilepsy project. Seizure. 2005; 14(7):514–20.
12. Perucca P, Gilliam FG. Adverse effects of antiepileptic drugs. Lancet Neurol [Internet]. 2012 [cited 2024]; 11(9):792–802. Available from: http://www.thelancet.com/article/S1474442212701539/fulltext
13. Xu Y, Nguyen D, Mohamed A, Carcel C, Li Q, Kutlubaev MA, et al. Frequency of a false positive diagnosis of epilepsy: A systematic review of observational studies. Seizure. 2016; 41:167–74.

14. Guerrini R. Epilepsy in children. Lancet. 2006/02/14. 2006; 367(9509):499–524.
15. Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. The Lancet. 2019; 393(10172):689–701.
16. Pitkänen A, Löscher W, Vezzani A, Becker AJ, Simonato M, Lukasiuk K, et al. Advances in the development of biomarkers for epilepsy. Lancet Neurol [Internet]. 2016 [cited 2022]; 15(8):843–56. Available from: https://pubmed.ncbi.nlm.nih.gov/27302363/
17. Van Donselaar CA, Stroink H, Arts WF. How confident are we of the diagnosis of epilepsy? Epilepsia [Internet]. 2006 [cited 2022]; 47 Suppl 1(SUPPL. 1):9–13. Available from: https://pubmed.ncbi.nlm.nih.gov/17044819/
18. Yew ANJ, Schraagen M, Otte WM, van Diessen E. Transforming epilepsy research: A systematic review on natural language processing applications. Epilepsia [Internet]. 2023 [cited 2023]; 64(2):292–305. Available from: https://onlinelibrary-wiley-com.proxy.library.uu.nl/doi/full/10.1111/epi.17474
19. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform [Internet]. 2013 [cited 2022]; 46(5):765–73. Available from: https://pubmed.ncbi.nlm.nih.gov/23810857/
20. Savova G, Pestian J, Connolly B, Miller T, Ni Y, Dexheimer JW, et al. Natural Language Processing: Applications in Pediatric Research. Translational Bioinformatics [Internet]. 2016 [cited 2022]; 10:231–50. Available from: https://link-springer-com.proxy.library.uu.nl/chapter/10.1007/978-981-10-1104-7_12
21. Buchlak QD, Esmaili N, Bennett C, Farrokhi F. Natural Language Processing Applications in the Clinical Neurosciences: A Machine Learning Augmented Systematic Review. Acta Neurochir Suppl [Internet]. 2022 [cited 2022]; 134:277–89. Available from: https://pubmed.ncbi.nlm.nih.gov/34862552/
22. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. ArXiv preprint arXiv:220607682 [Internet]. 2022 [cited 2023]; . Available from: https://openreview.net/forum?id=yzkSU5zdwD
23. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus. 2023; 15(5):e39305.
24. van Diessen E, van Amerongen RA, Zijlmans M, Otte WM. Potential merits and flaws of large language models in epilepsy care: A critical review. Epilepsia [Internet]. 2024 [cited 2024]; 65(4):873–86. Available from: https://onlinelibrary-wiley-com.utrechtuniversity.idm.oclc.org/doi/full/10.1111/epi.17907
25. Beaulieu-Jones BK, Villamar MF, Scordis P, Bartmann AP, Ali W, Wissel BD, et al. Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. Lancet Digit Health [Internet]. 2023 [cited 2023]; 5(12):e882–94. Available from: https://pubmed.ncbi.nlm.nih.gov/38000873/
26. Ford J, Pevy N, Grunewald R, Howell S, Reuber M. Can artificial intelligence diagnose seizures based on patients' descriptions? A study of GPT-4. medRxiv [Internet]. 2024 [cited 2024]; :2024.10.07.24314526. Available from: https://www.medrxiv.org/content/10.1101/2024.10.07.24314526v1
27. Xie K, Gallagher RS, Shinohara RT, Xie SX, Hill CE, Conrad EC, et al. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. Epilepsia [Internet]. 2023 [cited 2023]; 64(7):1900–9. Available from: https://pubmed.ncbi.nlm.nih.gov/37114472/

28. van Diessen E, Lamberink HJ, Otte WM, Doornebal N, Brouwer OF, Jansen FE, et al. A Prediction Model to Determine Childhood Epilepsy After 1 or More Paroxysmal Events. Pediatrics. 2018; 142(6).
29. Slinger G, Noorlag L, van Diessen E, Otte WM, Zijlmans M, Jansen FE, et al. Clinical characteristics and diagnoses of 1,213 children referred to a first seizure clinic. Epilepsia Open [Internet]. 2023 [cited 2024]; . Available from: https://onlinelibrary-wiley-com.proxy.library.uu.nl/doi/full/10.1002/epi4.12883
30. Fisher RS, Cross JH, French JA, Higurashi N, Hirsch E, Jansen FE, et al. Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. Epilepsia [Internet]. 2017 [cited 2017]; 58(4):522–30. Available from: http://doi.wiley.com/10.1111/epi.13670
31. Xu S. Bayesian Naïve Bayes classifiers to text classification. J Inf Sci. 2018; 44(1):48–59.
32. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. Journal of the American Medical Informatics Association [Internet]. 2011 [cited 2024]; 18(5):544–51. Available from: https://dx-doi-org.utrechtuniversity.idm.oclc.org/10.1136/amiajnl-2011-000464
33. Wikipedia the free encyclopedia. Document-term matrix. 2024.
34. Reimers N, Gurevych I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing [Internet]. 2020 [cited 2024]; :4512–25. Available from: https://github.com/facebookresearch/
35. Chen T, He T. xgboost: eXtreme Gradient Boosting. R package version 04-2. 2020; .
36. Wissel BD, Greiner HM, Glauser TA, Holland-Bouley KD, Mangano FT, Santel D, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. Epilepsia [Internet]. 2020 [cited 2022]; 61(1):39–48. Available from: https://pubmed.ncbi.nlm.nih.gov/31784992/
37. Connolly B, Matykiewicz P, Cohen KB, Standridge SM, Glauser TA, Dlugos DJ, et al. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. J Am Med Inform Assoc [Internet]. 2014 [cited 2022]; 21(5):866–70. Available from: https://pubmed.ncbi.nlm.nih.gov/24692393/
38. Keller AE, Ho J, Whitney R, Li SA, Williams AS, Pollanen MS, et al. Autopsy-reported cause of death in a population-based cohort of sudden unexpected death in epilepsy. Epilepsia [Internet]. 2021 [cited 2022]; 62(2):472–80. Available from: https://pubmed.ncbi.nlm.nih.gov/33400291/
39. Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, et al. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. J Biomed Inform [Internet]. 2014 [cited 2022]; 51:272–9. Available from: https://pubmed.ncbi.nlm.nih.gov/24973735/
40. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford D V., et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open [Internet]. 2019 [cited 2022]; 9(4). Available from: https://pubmed.ncbi.nlm.nih.gov/30940752/
41. Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach

to natural language processing. J Am Med Inform Assoc [Internet]. 2022 [cited 2023]; 29(5):873–81. Available from: https://pubmed.ncbi.nlm.nih.gov/35190834/

42. Meng Y, Elkaim L, Wang J, Liu J, Alotaibi NM, Ibrahim GM, et al. Social media in epilepsy: A quantitative and qualitative analysis. Epilepsy Behav [Internet]. 2017 [cited 2022]; 71(Pt A):79–84. Available from: https://pubmed.ncbi.nlm.nih.gov/28554148/
43. He K, Hong N, Lapalme-Remis S, Lan Y, Huang M, Li C, et al. Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups. Epilepsy Behav [Internet]. 2019 [cited 2022]; 94:65–71. Available from: https://pubmed.ncbi.nlm.nih.gov/30893617/
44. Galer PD, Parthasarathy S, Xian J, McKee JL, Ruggiero SM, Ganesan S, et al. Clinical signatures of genetic epilepsies precede diagnosis in electronic medical records of 32,000 individuals. Genetics in Medicine. 2024; 26(11):101211.
45. Lo Barco T, Garcelon N, Neuraz A, Nabbout R. Natural history of rare diseases using natural language processing of narrative unstructured electronic health records: The example of Dravet syndrome. Epilepsia [Internet]. 2024 [cited 2024]; 65(2):350–61. Available from: https://onlinelibrary-wiley-com.utrechtuniversity.idm.oclc.org/doi/full/10.1111/epi.17855
46. Pevy N, Christensen H, Walker T, Reuber M. Differentiating between epileptic and functional/dissociative seizures using semantic content analysis of transcripts of routine clinic consultations. Epilepsy & Behavior [Internet]. 2023 [cited 2024]; 143(6):109217. Available from: https://doi.org/10.1016/j.yebeh.2023.109217
47. Pevy N, Christensen H, Walker T, Reuber M. Feasibility of using an automated analysis of formulation effort in patients' spoken seizure descriptions in the differential diagnosis of epileptic and nonepileptic seizures. Seizure [Internet]. 2021 [cited 2022]; 91:141–5. Available from: https://pubmed.ncbi.nlm.nih.gov/34157636/
48. Padurariu C, Breaban ME. Dealing with Data Imbalance in Text Classification. Procedia Comput Sci. 2019; 159:736–45.
49. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMA: Open and Efficient Foundation Language Models. ArXiv preprint arXiv:230213971 [Internet]. 2023 [cited 2023]; . Available from: https://arxiv.org/abs/2302.13971v1
50. Köpf A, Kilcher Y, Von Rütte D, Anagnostidis S, Tam Z-R, Stevens K, et al. OpenAssistant Conversations-Democratizing Large Language Model Alignment. ArXiv preprint arXiv:230407327 [Internet]. 2023 [cited 2023]; . Available from: https://huggingface.co/OpenAssistant
51. Lamberink HJ, Otte WM, Geerts AT, Pavlovic M, Ramos-Lizana J, Marson AG, et al. Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: a systematic review and individual participant data meta-analysis. Lancet Neurol [Internet]. 2017 [cited 2017]; 16(7):523–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28483337
52. Stevelink R, Al-Toma D, Jansen FE, Lamberink HJ, Asadi-Pooya AA, Farazdaghi M, et al. Individualised prediction of drug resistance and seizure recurrence after medication withdrawal in people with juvenile myoclonic epilepsy: A systematic review and individual participant data meta-analysis. EClinicalMedicine [Internet]. 2022 [cited 2024]; 53. Available from: https://osf.io/b9zjc/

Appendix

Table A1. Baseline characteristics of the data.

Characteristics	Total, N (%)
Medical letters after the first consultation	
UMCU	1250 (80.1)
MZG	311 (19.9)
Sex	
Female	708 (45.4)
Male	853 (54.6)
Age	
Median	4,5
Mean	5,9
Highest age	17.8
Lowest age	0
Epilepsy diagnosis after first consultation	
Epilepsy	366 (23.5)
No epilepsy	795 (50.9)
Unclear	400 (25.6)
Epilepsy diagnosis after two years of follow-up	
Epilepsy	514 (32.9)
No epilepsy	958 (61.4)
Unclear	89 (5.7)
Epilepsy diagnosis after two years of follow-up from UMCU	
Epilepsy	413 (33.0)
No epilepsy	767 (61.4)
Unclear	70 (5.6)
Epilepsy diagnosis after two years of follow-up from MZG	
Epilepsy	101 (32.5)
No epilepsy	191 (61.4)
Unclear	19 (6.1)

Abbreviations: N = total number, UMCU = University Medical Center Utrecht, MZG = Martini Hospital Groningen.

Table A2. *A complete translation list of most prevalent features (Dutch-English)*

Features epilepsy group		**Features no epilepsy group**	
Dutch	***English***	***Dutch***	***English***
Spray	Spray	Tevreden	Satisfied
Kwijlde	Drooled	Ontstaan	Arise
1_spray	1_spray	Koortsstuip	Febrile seizure
Haar_mond	Her_mouth	Normaal_armen	Normal_arms
Insult_doorgemaakt	Experienced_seizure	Migraine	Migraine
Dubbele	Double	RR_na	Bloodpressure_after
Dubbele_tong	Slurred_speech	Moeder_en	Mother_and
Afgelopen_dagen	Last_days	Periode	Period
Smakt	Smack	Klachten	Complaints
Na_paar	After_few	Los	Loose
Insult_voorgeschiedenis	Seizure_history	Pakken	Take
Slijm	Slime	2-3	2-3
Komst_eerst	Visit_first	En_moeder	And_mother
Was_bleek	Was_pale	Leeftijd_van	Age_of
Niet_kon	Not_abled	Uren	Hours
Kon_hij	Abled_he	Hierover	Hereof
Draaien_naar	Turning_after	Een_andere	Another
Hij_lag	He_laid	Vanaf	From
Trof	Found	Anamnese_moeder	Anamnesis_mother
Moeder_hoorde	Mother_heard	Nek	Neck
Goed_geslapen	Slept_good	Met_wat	With_wath
Hemiparese	Hemiparesis	Geheel	Complete
Bij_familieleden	With_family members	Hard	Hard
Epileptisch_insult	Epileptic_seizure	Waarbij_ze	Where_they
Ouders_gescheiden	Parents_divorced	Peristaltiek	Peristalsis
Speeksel	Saliva	Verwijzer	Refferer
27	27	Geen_schokken	No_shocks
Eerste_epileptische	First_epileptic	Mediaan_motoriek	Median_motoric
Deze_aanval	This_attack	Symmetrisch	Symmetrical
Kwijlen	Drooling	Uitsteken	Protrude