Utrecht University

Faculty of Medicine
Academic year 2023-2024

Natural Language Processing for Early Diagnosis of Childhood Epilepsy

Jitse Loyens

Student Number: 6346065

Supervisor: Dr. E.G.A.L. van Diessen, Franciscus Gasthuis & Vlietland, Rotterdam; UMC Utrecht Brain Center

UMC Supervisor: Dr. W.M. Otte, UMC Utrecht Brain Center

Department: UMC Utrecht, Brain Center

Internship period: 20/05/2024 - 25/08/2024

List of abbreviations

Abbreviation	**Definition**
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
DFM	Document Feature Matrix
EEG	Electroencephalography
FSC	First Seizure Clinic
MRI	Magnetic Resonance Imaging
MZG	Martini Hospital Groningen
NLP	Natural Language Processing
RFE	Recursive Feature Elimination
TF-IDF	Term frequency-inverse document frequency
UMCU	University Medical Center Utrecht
URL	Uniform Resource Locator

Abstract

Background: Accurate and timely diagnosis is crucial in epilepsy treatment. Diagnostic delay in epilepsy results in unnecessary risk exposure to psychosocial distress, morbidity, or mortality. Language is an indispensable source of information for diagnosing epilepsy. Natural language processing, a branch of artificial intelligence, analyses language to extract information and identify patterns. This study assessed the diagnostic value of natural language processing to facilitate the early diagnosis of childhood epilepsy.

Methods: A dataset of 1561 letters from first consultations was available from the University Medical Center Utrecht and Martini Hospital Groningen. Natural language processing was applied to analyse textual data and classify the letters as either 'epilepsy' or 'no epilepsy'. The Naïve Bayes model was employed for text classification. Data was divided into training and test sets to evaluate performance and generalisability. Training sets identified predictive features, consisting of keywords indicative of 'epilepsy' or 'no epilepsy'. The model's output was compared to the clinician's final diagnosis (gold standard).

Results: Model accuracy ranges from 0.66 to 0.68. Balanced accuracy varies from 0.67 to 0.72 for ‘epilepsy’ and 0.68 to 0.73 for ‘no epilepsy’. F1 score varies from 0.50 to 0.57 for 'epilepsy' and 0.76 to 0.80 for 'no epilepsy'. AUROC varies from 0.74 to 0.78 for ‘epilepsy’ and 0.73 to 0.77 for ‘no epilepsy’. AUPRC varies from 0.52 to 0.63 for ‘epilepsy’ and 0.79 to 0.81 for ‘no epilepsy’.

Conclusion: All models demonstrated moderate to good performance, with better performance in diagnosing ‘no epilepsy’. Improvements are required to enhance accuracy and generalisability.

Introduction

Epilepsy is one of the most common neurological disorders and one of the leading neurological causes of morbidity and mortality.[1,2,3] Epilepsy affects 50-70 million people worldwide.[4,5] In the Netherlands, approximately 211,800 people suffer from epilepsy.[6] In Belgium, approximately 1 in 200 people are affected by epilepsy, which corresponds to around 60,000 individuals in Flanders.[7,8,9] Individuals under the age of 16 constitute 40% of the population with epilepsy.[9]

Epilepsy is defined as "two unprovoked seizures occurring more than 24 hours apart, a single unprovoked seizure with a high recurrence risk (>60% over the next 10 years), or a diagnosis of an epilepsy syndrome."[10] Epilepsy is characterised by a predisposition to generate epileptic seizures, which result from abnormal electrical discharges in the brain.[10,11,12,13]

Epilepsy is associated with significant cognitive, psychiatric, and physical comorbidities.[2,14] More than half of epilepsy patients experience additional medical conditions.[15] Epilepsy also has harmful effects on social and psychological well-being including stigma and social isolation, anxiety, limitations on daily activities, cognitive dysfunction, and issues related to work, school, and relationships.[13,16,17,18] According to the Global Burden of Disease study (2010), severe epilepsy ranked fourth among 220 health conditions regarding disability weight.[13,19] Epilepsy increasingly contributes to global disability-adjusted life years and mortality.[5,20,21,22] Epilepsy patients are at an increased risk of death, which may be related directly to seizures (through status epilepticus or 'sudden unexpected death in epilepsy') or indirectly to associated non-seizure factors (through injuries, drowning, or aspiration pneumonia).[15,16,23]

Diagnosing epilepsy can be complex and demands considerable time and effort.[15,24] Due to its polymorphic nature, epilepsy can manifest in various ways and has numerous mimics.[15] Consequently, seizures are often incorrectly diagnosed and under-detected.[25] Diagnostic delay is an increasingly recognised issue and may lead to undesired health conditions.[22] Various studies have reported a significant diagnostic delay in epilepsy. Parviainen et al. demonstrated a median delay of 12 months for new-onset focal epilepsy.[26] Similarly, 41% of children who experienced their first seizure before the age of three had diagnostic delays exceeding a month, with 13% extending beyond a year.[27] Slinger et al. reported that 8.7% of children at a Dutch tertiary children's hospital received their final diagnosis after 12 months.[28] Moreover, nearly half of the patients assessed for initial seizures were experiencing recurrent, undiagnosed seizures at the time of evaluation.[25,29] Generally, diagnostic time is brief for clearly diagnosable epilepsy but can extend beyond a year for complex or ambiguous cases. Such diagnostic delays can result in impaired cognition, reduced quality of life, physical injuries, and increased risk of mortality.[22,29,30,31,32,33,34] However, a false-positive diagnosis may result in the unnecessary administration of antiseizure medication, which can cause adverse effects, including neuropsychiatric symptoms.[35,36,37] False-positive diagnoses are estimated to occur in up to 25% of patients.[38,39] Xu et al. reported a wide range in the frequency of false-positive epilepsy misdiagnosis, from 2% to 71%, with syncope and psychogenic non-epileptic paroxysmal events being the most common mimics.[35] Nonetheless, an estimated 70% of epilepsy patients could be seizure-free with appropriate diagnosis

and treatment.[4] Furthermore, a delayed or misdiagnosis incurs financial costs from unnecessary medical assessments and inappropriate treatments.[22]

Language is widely recognised as an indispensable source of information for diagnosing epilepsy, evaluating treatment and managing patient care.[40] Clinicians take history and distil relevant clinical information from a patient's narrative.[40] Language serves as a rich and versatile medium for obtaining deep insight into the patient's condition, essential for a holistic approach to epilepsy care. Despite advancements in ancillary investigations (EEG, MRI, and genetic testing), clinical information from electronic health records remains indispensable for diagnosing and monitoring epilepsy.[15,41,42] However, this wealth of information is often stored and collected in patient records in an unstructured manner, limiting its optimal utilisation in clinical decision-making.[1,43]

The advent of natural language processing (NLP) to systematically process unstructured textual data presents an unprecedented opportunity to utilise this information source for clinical purposes.[40] NLP is a form of artificial intelligence specialising in the computational analysis of spoken and written language to identify patterns and trends and extract relevant information.[40] This involves converting unstructured text into a structured format. Computational algorithms are then applied to process and analyse these structured features, enabling the retrieval of the desired information that supports diagnosis or decision-making.[1] In the medical domain, NLP can be applied to electronic patient records, clinical notes, letters, patient messages, scientific articles, audio recordings, or medical guidelines, making a previously unexploited source of information available for clinical purposes.[40,44,45] NLP has therefore acquired increasing popularity across various medical subfields. It has been demonstrated to be helpful for the early detection and classification of diverse health conditions and child abuse.[46,47,48,49,50]

In epilepsy research, there is an increasing tendency towards applying NLP for patient identification, risk stratification, and prediction.[51,52,53] In clinical settings, NLP can contribute to the early detection of medical conditions, thereby reducing the time to diagnosis and treatment.[1] NLP algorithms are able to identify implicit textual patterns predictive of a medical condition.[1] Despite the potential of NLP, the application of NLP for the early diagnosis of epilepsy based on medical documentation has not yet been explored in the literature. Therefore, this study aims to assess the diagnostic value of applying NLP to medical letters from the first consultation to facilitate the early diagnosis of childhood epilepsy.

Methodology

Dataset

The dataset consists of 1561 medical patient letters, with 1250 originating from University Medical Center Utrecht (UMCU) and 311 from Martini Hospital Groningen (MZG). Data were retrospectively collected from children (age < 18 years) who were referred to the First Seizure Clinic (FSC) between 2008 and May 2022. Data were originally collected for previously published studies focusing on a prediction model development for childhood epilepsy and the clinical characteristics and diagnoses of children referred to an FSC.[28,54] The letters were written by various paediatric neurologists. Patient characteristics included sex and age at first seizure. For each patient, the initial diagnosis (i.e., the diagnosis established at the end of the first FSC consultation) and the final diagnosis (i.e., the diagnosis reached through consensus among doctors and/or ancillary investigations at the latest follow-up, recorded within a two-year period) were added to the dataset. Follow-up occurred for children with inconclusive diagnoses at the first consultation and for those initially diagnosed with epilepsy. The initial and final diagnoses encompass the groups 'epilepsy', 'no epilepsy', and 'unclear' (refer to Figure 1 for the graphic representation of the diagnoses). Epilepsy diagnoses are established according to the International League Against Epilepsy definition of epilepsy.[10] The initial diagnosis was considered 'unclear' if ancillary investigations were assessed as necessary to confirm or reject the epilepsy diagnosis.[28] The final diagnosis was classified as 'unclear' if, despite further investigations, it remained uncertain whether the events were indeed related to epilepsy.[28]

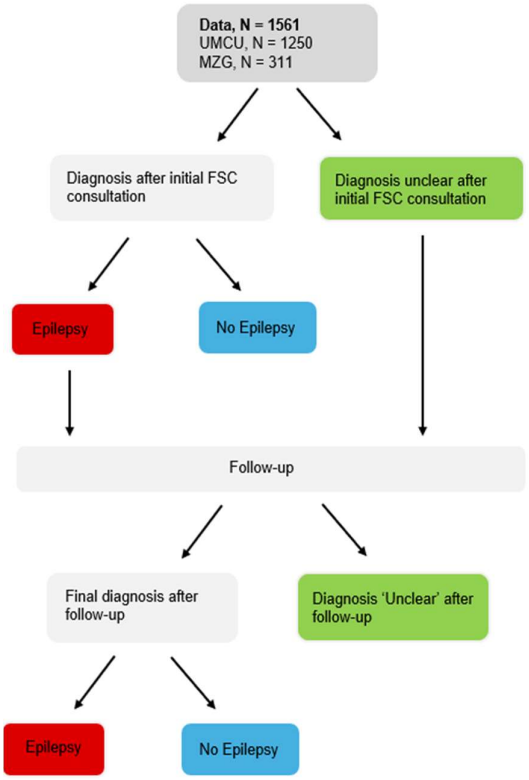

Figure 1. Flowchart illustrating the diagnostic pathway for children referred to the FSC. The flowchart outlines the process from the first FSC consultation to the final diagnosis, including follow-up procedures. The diagnoses are categorised as 'epilepsy', 'no epilepsy', or 'unclear'.

Study design

A retrospective analysis was conducted on letters to assess the clinical application of NLP for the early diagnosis of childhood epilepsy. This was achieved through text classification, specifically by training classification models based on textual features and predicting the class of new texts. A classification model was developed on a training set and subsequently tested on a test set. The final diagnosis was considered the gold standard for evaluating the performance of the NLP model. Different analyses were performed. In Analysis 1A, data were randomly divided into a training and a test set, with a respective

ratio of 70% (1094 letters) and 30% (467 letters). To ensure a representative distribution of the final diagnosis in both sets, stratification was applied based on the final diagnosis groups. Analysis 1B utilised the training set previously established in Analysis 1A. The test set comprised all 400 cases from the 'unclear' group, where the diagnosis was indeterminate after the first consultation (initial FSC diagnosis). This analysis attempts to determine whether the model can accurately classify patients with an unclear diagnosis at the first consultation as either having epilepsy or not. Analysis 2 assessed the generalisability (i.e. external validity). Therefore, the training set comprised data from UMCU (1250 letters) and the test set comprised data from MZG (311 letters).

NLP workflow

The NLP workflow consists of three main phases: data preprocessing, data analysis, and classification. This workflow is illustrated in Figure 2.

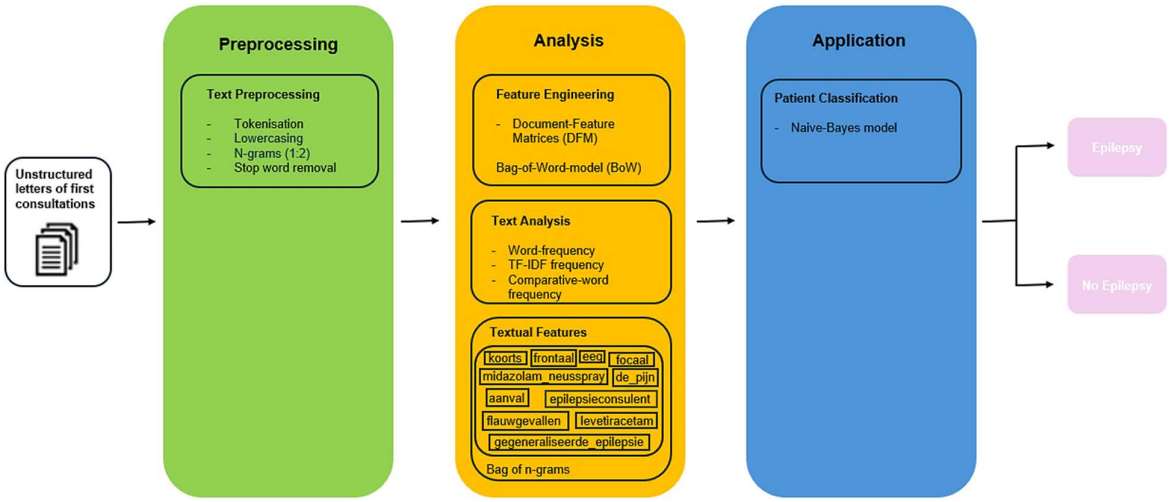

Figure 2. NLP workflow for classifying 'epilepsy' or 'no epilepsy' diagnosis based on unstructured letters from the first consultations. The process consists of three main stages: preprocessing, analysis, and application.

Data preprocessing

Data preprocessing encompasses several key steps including corpus creation, tokenisation, data cleaning, lowercasing, n-gram generation, and stop word removal. Prior to data preprocessing, the ancillary investigations, conclusion, treatment plan, and considerations were omitted from all letters to ensure the focus remained on the available information during the first consultation. This approach also prevents the model from depending on or being biased by conclusions or treatment plans, thereby reducing interpretative bias. Creating a corpus involves collecting and organising a substantial amount of textual data in a structured manner to facilitate systematic analysis and processing. The text was then divided into tokens (i.e., words) through tokenisation. Undesired characters, such as punctuation marks, symbols, URLs, and separators, were omitted. Lowercasing converted all characters in the text to lowercase letters, ensuring consistency across the tokens. Afterwards, n-grams were generated. 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”) and bigrams (pairs of consecutive words such as “geen_trekkingen”). 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” may be retained while “geen” and “koorts” 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. A list of the removed stop words is provided in Table A1, in Appendix 1.

Data analysis

A document-feature matrix (DFM) was created. A DFM employs a matrix representation of the data, which allows for the structured analysis of text data by representing documents (letters) as rows and features (n-grams) as columns in the matrix.[55] The values in the matrix represent the frequency of a feature in a letter.[55] In this manner, a Bag-of-Words (BoW) model was created, where each text is represented as a collection of words, disregarding the order in which they appear.[56]

The following analyses were performed: a word-frequency analysis, a term frequency-inverse document frequency (TF-IDF) analysis, and a comparative word-frequency analysis. Word-frequency analysis was conducted to identify the most common features within the dataset. TF-IDF analysis was conducted to identify the most significant features within the dataset based on their TF-IDF scores. TF-IDF was applied to weigh features based on their frequency in individual letters.TF-IDF reduces the influence of frequently occurring features and emphasises those features that are more informative for classification. Comparative word-frequency analysis was performed to identify features that are significantly more frequent in one group compared to another.

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 minimise overfitting and optimise 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 minimising noise, thereby enhancing the robustness and generalisability 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.

Classification

The Naïve Bayes classification was implemented, with the DFM as input. The Naïve Bayes classifier was selected due to its simplicity and effectiveness in text classification. This probabilistic model applies Bayes' theorem “with strong (naive) independence assumptions between features”.[57] As a hyperparameter for the Naive Bayes model, the smoothing parameter (α) was added to prevent zero

probabilities. To address class imbalance, class weights were applied, reducing the impact of the predominant group ‘no epilepsy’ and improving performance in underrepresented groups.

Performance evaluation

A confusion matrix evaluated the classification model’s performance by comparing actual and predicted classifications employing decision statistics from contingency tables. Performance metrics include accuracy, recall (sensitivity), precision (positive predictive value), AUROC (Area Under the Receiver Operating Characteristic Curve), AUPRC (Area Under the Precision-Recall Curve), and F1 score (i.e., “the harmonic mean of the precision and recall”).[58] All analyses were performed with R software, version 4.4.0.[59]

Results

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 (95%CI: 17.3-17.8), while the minimum age was 0 years (95%CI: 0.0-0.0). 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 in Table A2, in Appendix 2.

Most important features

The most important features, including all figures, are presented in Appendix 3.

Classification model performance

Analyse 1A

The training set demonstrated an overall accuracy of 0.77 (95%CI: 0.7388-0.7899). 'No epilepsy' demonstrated higher balanced accuracy, recall, precision, F1-scores, AUROC and AUPRC values compared to 'epilepsy'. The test set demonstrated an overall accuracy of 0.66 (95%CI: 0.6102-0.6983). In contrast, ‘epilepsy’ demonstrated higher balanced accuracy and AUROC value compared to ‘no epilepsy’. However, ‘no epilepsy’ demonstrated higher recall, precision, F1-score, and AUPRC values compared to ‘epilepsy’. Table 1 provides the performance metrics of the NLP model on the training and test sets.

Table 1. Class-specific performance metrics for epilepsy diagnosis: training and test sets, analysis 1A.

Analysis set	Class	Balanced accuracy	F1-score	Recall	Precision	AUROC	AUPRC
Training set	Epilepsy	0.76	0.67	0.70	0.64	0.84	0.75
	No Epilepsy	0.79	0.84	0.83	0.84	0.85	0.89
Test set	Epilepsy	0.70	0.57	0.62	0.53	0.78	0.63
	No Epilepsy	0.68	0.76	0.74	0.78	0.73	0.80

The table presents the balanced accuracy, F1-score, recall (sensitivity), precision (positive predictive value), AUROC, and AUPRC for the 'epilepsy' and 'no epilepsy' classes. These metrics provide a comprehensive overview of the NLP model's performance in classifying the diagnosis groups within the training and test sets.

Analyse 1B

The test set demonstrated an overall accuracy of 0.66 (95%CI: 0.6087-0.7039). Table 2 provides the performance metrics of the NLP model on the test set. 'Epilepsy' and 'no epilepsy' exhibited similar AUROC values, but 'no epilepsy' achieved higher balanced accuracy, recall, precision, and F1 score compared to 'epilepsy'. The lower F1-score for 'epilepsy' emphasises the difficulties associated with accurately classifying this condition.

Table 2. Class-specific performance metrics for epilepsy diagnosis: test set, analysis 1B.

Analysis set	Class	Balanced accuracy	F1-score	Recall	Precision	AUROC	AUPRC
Test set	Epilepsy	0.67	0.50	0.49	0.50	0.77	0.52
	No Epilepsy	0.72	0.76	0.72	0.79	0.77	0.79

The table presents the balanced accuracy, F1-score, recall (sensitivity), precision (positive predictive value), AUROC, and AUPRC for the 'epilepsy' and 'no epilepsy' classes. These metrics provide a comprehensive overview of the NLP model's performance in classifying the diagnosis groups within the training and test sets.

Analyse 2

The training set demonstrated an overall accuracy of 0.74 (95%CI: 0.7164-0.7657). Balanced accuracy was consistent between the two classes. However, recall, precision, F1-score, AUROC and AUPRC values were slightly lower compared to analysis 1, indicating a potential reduction in the model's robustness in this analysis. The test set demonstrated an overall accuracy of 0.68 (95%CI: 0.6267-0.7331). Balanced accuracy was also relatively consistent between the two classes but 'no epilepsy' demonstrated superior performance metrics across all measures. Table 3 provides the performance metrics of the NLP model on the training and test sets.

Table 3. Class-specific performance metrics for epilepsy diagnosis: training and test sets, analysis 2.

Analysis set	Class	Balanced accuracy	F1-score	Recall	Precision	AUROC	AUPRC
Training set	Epilepsy	0.76	0.67	0.69	0.64	0.83	0.73
	No Epilepsy	0.76	0.81	0.82	0.81	0.83	0.87

Test set	Epilepsy	0.72	0.52	0.68	0.43	0.74	0.58
	No Epilepsy	0.73	0.80	0.76	0.84	0.77	0.81

The table presents the balanced accuracy, F1-score, recall (sensitivity), precision (positive predictive value), AUROC, and AUPRC for the 'epilepsy' and 'no epilepsy' classes. These metrics provide a comprehensive overview of the NLP model's performance in classifying the diagnosis groups within the training and test sets.

Confusion matrices, AUROC and AUPRC curves for each analysis are provided in Appendix 4 (Figures A12-A14 for Analysis 1A, A15-A17 for Analysis 1B, and A18-A20 for Analysis 2).

Discussion

Principal findings

This study aimed to assess the diagnostic value of applying NLP to medical letters from the first consultation to facilitate the early diagnosis of childhood epilepsy. The results demonstrate that the NLP model achieves superior performance in classifying cases as ‘no epilepsy’ compared to those with epilepsy. This is particularly evident from the consistently higher F1-scores, precision, and AUPRC values observed for 'no epilepsy'. However, the variability in performance, especially across the test sets, emphasises the challenges the model encounters in generalising to new data. In a clinical context, these findings indicate that the model is proficient at diagnosing no epilepsy, essential for minimising unnecessary diagnostic procedures and focusing on appropriate treatments. Nevertheless, the variability in performance illustrates the difficulties in achieving consistent accuracy. Additionally, the model is less effective at diagnosing epilepsy, which may impact the model's clinical utility. These results emphasise the necessity for continuous refinement and validation of NLP tools to improve their diagnostic accuracy and reliability, thereby facilitating more accurate and timely diagnoses of childhood epilepsy.

Comparison with prior work

Comparing results with prior studies contextualises the NLP model performance in diagnosing childhood epilepsy. Several studies provide valuable insights into similar NLP applications. Chase et al. investigated the application of NLP for the early recognition of multiple sclerosis by analysing electronic health records. The Naïve Bayes model was employed. They reported an AUROC of 0.94, with a sensitivity of 81% and a specificity of 87% for classifying diagnosed multiple sclerosis patients. Additionally, the model identified 40% of multiple sclerosis patients before the official diagnosis.[46] Fernandes et al. investigated automated electronic health record phenotyping for identifying patients with epilepsy. They employed logistic regression and extreme gradient boosting models. The model achieved a macro average AUROC and AUPRC of 1.00 on the test set.[51] Shahi et al. employed a Bidirectional Encoder Representations from Transformers (BERT) model to detect child physical abuse. The model had an accuracy of 86%, an F1 score of 0.86, and an AUROC of 0.86.[50] Importantly, these studies involved different populations, feature selection methods, and models. For instance, Chase et al. and Fernandes et al. focused on adult populations.[46,51] Chase et al. and Shahi et al. applied NLP to facilitate the early detection of specific conditions.[46,50] In contrast, Fernandes et al. aimed to identify patients with a confirmed diagnosis of epilepsy.[51] This objective differs significantly, although there are similarities in the condition-specific features. However, Fernandes et al. also incorporated features from ancillary investigations (EEG and MRI) and treatment plans into the model. They utilised 286 features, with feature selection performed by L1 regularisation (Lasso regression).[51] Chase et al. did not specify the number of features but relied on predefined word lists relevant to multiple sclerosis, including commonly associated medical terminology and symptoms.[46] Shahi et al. employed a BERT model, which does not require feature selection.[50] Instead, BERT utilises a transformer architecture to learn contextual word relationships and automatically determine the most important features. Fernandes et

al. and Shahi et al. employed more advanced models, resulting in improved performance metrics compared to the Naïve Bayes model.[50,51]

Strengths and limitations

This study is pioneering in its application of NLP for the early diagnosis of childhood epilepsy. While NLP has been explored for diagnosing other conditions and achieving various objectives, this study specifically addresses the early diagnosis of epilepsy in a paediatric population. This approach bridges a significant gap in the existing research. The study benefits from a diverse dataset comprising 1561 medical patient letters from two hospitals. The substantial size and diversity of the dataset contribute to the robustness of the study's findings and enhance the generalizability of the results across various clinical settings. The study also utilised well-defined diagnostic categories, including 'epilepsy', 'no epilepsy', and 'unclear'. The model's performance was evaluated through various metrics, thoroughly assessing its effectiveness in distinguishing between the diagnostic groups. ROC curves demonstrate superior performance compared to other metrics. The ROC curve and AUROC provide a comprehensive assessment of model performance by analysing the trade-off between true positives and false positives across all possible thresholds. This makes them particularly valuable for evaluating a model's overall discriminatory capability.

Performance was significantly higher on the training sets 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 generalisability to new data.[60] This can result from excessive noise, an excessive number of features, irrelevant features, or insufficient training data. Despite applying class weights, the model may experience difficulties correctly integrating these weights into the learning algorithm. Additionally, imbalanced data can still challenge the model's ability to accurately learn the complexity of the minority class. Models such as Naïve Bayes have an inherent bias towards the majority class, as they assume conditional independence between features. Potential solutions include dataset balancing through oversampling the minority class or undersampling the majority class. In cases of uneven class distributions, overall accuracy can be misleading as it reflects the percentage of correctly classified cases without accounting for class distribution. In such situations, balanced accuracy, which considers performance across both classes by calculating the average recall, provides a more equitable assessment. Furthermore, the Naive Bayes model is relatively simplistic and limited in its capacity to learn complex relationships. More advanced models, such as BERT, provide a viable alternative. Moreover, the Naive Bayes model does not consider word order. For instance, in a list such as "geen koorts, trekkingen, tongbeet, bewustzijnsverlies", it may only recognise "geen_koorts". Similarly, "geen_trekkingen" could be misinterpreted as "trekkingen". N-grams do not always effectively recognise negations either. For example, in the sentence "het is geen insult", "geen_insult" might be incorrectly processed as "is_geen". This can be addressed by employing transformer models or expanding n-grams to include sequences such as four consecutive words. Additionally, confounding factors like typographical errors, abbreviations, double negations, and letters written by multiple authors can adversely affect the classification process. Furthermore, RFE was applied to a subset of the top 8000 features 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. Moreover, lemmatisation was not implemented during preprocessing for simplicity. Lemmatisation reduces words to their lemma, the base form of the word. Instead of working with various inflected forms of a word, such as plural forms or verb conjugations, all these variations are consolidated into the base form. For example, the words “smakt”, “smakte”, and “smakken” will all be reduced to “smakken”. Lastly, in Analysis 1B, the same training set was utilised as in Analysis 1A, which could lead to an overlap between the training and test sets of Analysis 1B. This overlap may result in an overestimation of precision in Analysis 1B.

Future directions

This study is retrospective, which is less optimal for evaluating diagnostic methods. Prospective studies provide better control over variables, reduce data noise, and allow for the direct capture of real-time language, rather than relying on historical records that may be incomplete or biased. Collecting data contemporaneously minimises data contamination and provides a more reliable assessment of NLP models. Future research should incorporate a prospective design to explore the clinical applicability of NLP. Several strategies should be considered to enhance model performance. Preprocessing could be improved by incorporating lemmatisation and expanding n-grams. Refined feature selection, applied to the entire dataset, can help focus on the most relevant features while excluding irrelevant ones. Implementing model validation methods, such as cross-validation, is important for evaluating performance and mitigating overfitting. Addressing data imbalance is essential for improving classification performance across all groups. Increasing the dataset size will provide a more robust foundation for training, thereby enhancing model reliability. Exploring advanced models, such as deep learning approaches, may yield improvements in predictive accuracy. Moreover, future research should assess the model's generalisability by evaluating its performance across datasets from multiple hospitals to ensure broader applicability and effectiveness. Additionally, the model could be integrated with other medical data sources, including notes from electronic health records, patient questionnaires, and ancillary investigation reports. Combining the classification model with a predictive epilepsy model, such as the one developed by Van Diessen et al., could also be explored.[54] Another future direction is real-time monitoring, where a clinician receives real-time feedback from the NLP model regarding the history taking while documenting clinical information from the patient.[1] Once the clinician completes the data entry, the NLP model promptly generates a probability of epilepsy based on the clinical notes.[1] This can improve treatment strategies and patient counselling. This study focused on written text. However, NLP can also be applied to spoken text. For instance, Pevy et al. investigated the feasibility of employing automated analysis of formulation effort in patients’ spoken seizure descriptions for differential diagnosis of epileptic and non-epileptic seizures.[61] The study demonstrated promise in distinguishing between epileptic and non-epileptic seizures by analysing spoken language.[61] Analysing spoken language allows for incorporating emotional tone and sentiment, providing deeper insights into the patient’s experience. Furthermore, ethical considerations play a role in applying NLP to medical data, particularly regarding patient privacy and data protection. Strict adherence to data protection regulations is essential to ensure confidentiality and security.

Conclusion

The application of NLP for the early diagnosis of epilepsy demonstrates moderate to good diagnostic value, with better model performance in detecting ‘no epilepsy’ cases. Despite the potential of NLP, significant improvements are required to improve the accuracy and generalisability of the classification model. This study provides new perspectives on integrating NLP into medical diagnostics and establishes a foundation for further research and development in this domain.

References

1. Yew ANJ, Schraagen M, Otte WM, van Diessen E. Transforming epilepsy research: A systematic review on natural language processing applications. Epilepsia. 2023 Feb 1;64(2):292–305.

2. Keezer MR, Sisodiya SM, Sander JW. Comorbidities of epilepsy: Current concepts and future perspectives. Lancet Neurol. 2016 Jan 1;15(1):106–15.

3. Kerr WT, McFarlane KN. Machine Learning and Artificial Intelligence Applications to Epilepsy: a Review for the Practicing Epileptologist. Vol. 23, Current Neurology and Neuroscience Reports. Springer; 2023. p. 869–79.

4. World Health Organization. Epilepsy overview. [Internet]. Available from: https://www.who.int/news-room/fact-sheets/detail/epilepsy. [Accessed 6th June 2024].

5. Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR. Estimation of the burden of active and life-time epilepsy: A meta-analytic approach. Epilepsia. 2010 May;51(5):883–90.

6. VZinfo. Epilepsie | Leeftijd en geslacht. [Internet]. Available from: : https://www.vzinfo.nl/epilepsie/leeftijd-en-geslacht#:~:text=Naar%20schatting%2062.500%20mensen%20met,32.800%20mannen%20en%2029.700%20vrouwen. [Accessed 6th June 2024].

7. Gezondheid.be. Epilepsie treft in Vlaanderen zo'n 60.000 mensen [Internet]. Available from: https://www.gezondheid.be/artikel/epilepsie/epilepsie-treft-in-vlaanderen-zon-60000-mensen-29809. [Accessed 6th June 2024].

8. UZgent. Epilepsie: diagnose en aanpak van moeilijk behandelbare vormen [Internet]. Available from: https://www.uzgent.be/sites/default/files/documents/epilepsie.pdf. [Accessed 6th June 2024].

9. UZA. Epilepsie [Internet]. Available from: https://www.uza.be/behandeling/epilepsie#:~:text=Epilepsie%20is%20na%20migraine%20de,elektrisch%20verschijnsel%20in%20de%20hersenen. [Accessed 6th June 2024].

10. Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE Official Report: A practical clinical definition of epilepsy. Epilepsia. 2014;55(4):475–82.

11. Fisher RS, Van Emde Boas W, Blume W, Elger C, Genton P, Lee P, et al. Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia. 2005 Apr;46(4):470–2.

12. Mora S, Turrisi R, Chiarella L, Consales A, Tassi L, Mai R, et al. NLP-based tools for localization of the epileptogenic zone in patients with drug-resistant focal epilepsy. Sci Rep. 2024 Dec 1;14(1).

13. Perucca P, Scheffer IE, Kiley M. The management of epilepsy in children and adults. Medical Journal of Australia. 2018 Mar 19;208(5):226–33.

14. Yuen AWC, Keezer MR, Sander JW. Epilepsy is a neurological and a systemic disorder. Epilepsy and Behavior. 2018 Jan 1;78:57–61.

15. Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. The Lancet. 2019 Feb 16;393(10172):689–701.

16. Foster E CPLDAZOTKP. First seizure presentations in adults: beyond assessment and treatment. . J Neurol Neurosurg Psychiatry . 2019 Sep;90(9):1039–45.

17. Steiger BK, Jokeit H. Why epilepsy challenges social life. Vol. 44, Seizure. W.B. Saunders Ltd; 2017. p. 194–8.

18. Velissaris SL, Wilson SJ, Newton MR, Berkovic SF, Saling MM. Cognitive complaints after a first seizure in adulthood: Influence of psychological adjustment. Epilepsia. 2009 May;50(5):1012–21.

19. Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, et al. Common values in assessing health outcomes from disease and injury: Disability weights measurement study for the Global Burden of Disease Study 2010. The Lancet. 2012;380(9859):2129–43.

20. Greenlund SF, Croft JB, Kobau R. Epilepsy by the Numbers: Epilepsy deaths by age, race/ethnicity, and gender in the United States significantly increased from 2005 to 2014. Epilepsy and Behavior. 2017 Apr 1;69:28–30.

21. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet [Internet]. 2018 Nov;392(10159):1736–88. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673618322037

22. Pellinen J, French J, Knupp KG. Diagnostic Delay in Epilepsy: the Scope of the Problem. Curr Neurol Neurosci Rep. 2021 Dec 1;21(12).

23. Devinsky O, Spruill T, Thurman D, Friedman D. VIEWS & REVIEWS Recognizing and preventing epilepsy-related mortality A call for action. 2015.

24. Brown RJ TM. Dissociative psychopathology, non-epileptic seizures, and neurology. 2000;285–9. Available from: www.jnnp.com

25. Jackson A, Teo L, Seneviratne U. Challenges in the first seizure clinic for adult patients with epilepsy. Epileptic Disorders. 2016 Sep 1;18(3):305–14.

26. Parviainen L, Kälviäinen R, Jutila L. Impact of diagnostic delay on seizure outcome in newly diagnosed focal epilepsy. Epilepsia Open. 2020 Dec 1;5(4):605–10.

27. Berg AT, Loddenkemper T, Baca CB. Diagnostic delays in children with early onset epilepsy: Impact, reasons, and opportunities to improve care. Epilepsia. 2014 Jan;55(1):123–32.

28. 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. 2024 Apr 1;9(2):548–57.

29. 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 May 7;92(19):E2197–208.

30. Lee SA, Kim MJ, Lee HW, Heo K, Shin DJ, Song HK, et al. The effect of recurrent seizures on cognitive, behavioral, and quality-of-life outcomes after 12months of monotherapy in adults with newly diagnosed or previously untreated partial epilepsy. Epilepsy and Behavior. 2015 Dec 1;53:202–8.

31. 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. 2015 Oct 1;56(10):1534–41.

32. Pellinen J, Tafuro E, Yang A, Price D, Friedman D, Holmes M, et al. Focal nonmotor versus motor seizures: The impact on diagnostic delay in focal epilepsy. Epilepsia. 2020 Dec 1;61(12):2643–52.

33. Novy J, Belluzzo M, Caboclo LO, Catarino CB, Yogarajah M, Martinian L, et al. The lifelong course of chronic epilepsy: The Chalfont experience. Brain. 2013;136(10):3187–99.

34. Nevalainen O, Ansakorpi H, Simola M, Raitanen J, Isojärvi J, Artama M, et al. VIEWS & REVIEWS Epilepsy-related clinical characteristics and mortality A systematic review and meta-analysis. 2014;

35. 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. Vol. 41, Seizure. W.B. Saunders Ltd; 2016. p. 167–74.

36. Perucca P, Gilliam FG. Adverse effects of antiepileptic drugs. Vol. 11, The Lancet Neurology. 2012. p. 792–802.

37. Gaitatzis A, Sander JW. The long-term safety of antiepileptic drugs. Vol. 27, CNS Drugs. 2013. p. 435–55.

38. Scheepers B, Clough P, Pickles C. The misdiagnosis of epilepsy: findings of a population study*. Vol. 7, Seizure. 1996.

39. Leach JP, Lauder R, Nicolson A, Smith DF. Epilepsy in the UK: Misdiagnosis, mistreatment, and undertreatment? Seizure. 2005 Oct;14(7):514–20.

40. 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. 2024 Apr 1;65(4):873–86.

41. Bankstahl JP, Pitkänen A, Löscher W, Vezzani A, Becker AJ, Simonato M, et al. Advances in the development of biomarkers for epilepsy [Internet]. Vol. 15, Review Lancet Neurol. 2016. Available from: www.thelancet.com/neurology

42. Van Donselaar CA, Stroink H, Arts WF. How confident are we of the diagnosis of epilepsy? Vol. 47, Epilepsia. 2006. p. 9–13.

43. Sheikhalishahi S RMRDJLARFO V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform. 2019 Apr;7(2).

44. Takale DG. A Study of Natural Language Processing in Healthcare Industries. 2024; Available from: https://doi.org/10.48001/JoWACS.2024.221-6

45. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. Vol. 77, Journal of Biomedical Informatics. Academic Press Inc.; 2018. p. 34–49.

46. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak. 2017 Feb 28;17(1):24.

47. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg. 2017 Jun 1;65(6):1753–61.

48. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, et al. Building a Natural Language Processing Tool to Identify Patients with High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. In: Academic Emergency Medicine. Blackwell Publishing Inc.; 2016. p. 628–36.

49. Tibbo ME, Wyles CC, Fu S, Sohn S, Lewallen DG, Berry DJ, et al. Use of Natural Language Processing Tools to Identify and Classify Periprosthetic Femur Fractures. Journal of Arthroplasty. 2019 Oct 1;34(10):2216–9.

50. Shahi N, Shahi AK, Phillips R, Shirek G, Lindberg DM, Moulton SL. Using deep learning and natural language processing models to detect child physical abuse. J Pediatr Surg. 2021 Dec 1;56(12):2326–32.

51. Fernandes M, Cardall A, Jing J, Ge W, Moura LMVR, Jacobs C, et al. Identification of patients with epilepsy using automated electronic health records phenotyping. Epilepsia. 2023 Jun 1;64(6):1472–81.

52. 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. 2023 Dec 1;5(12):e882–94.

53. Wissel BD, Greiner HM, Glauser TA, Pestian JP, Ficker DM, Cavitt JL, et al. Early Identification of Candidates for Epilepsy Surgery: A Multicenter, Machine Learning, Prospective Validation Study. Neurology. 2024 Feb 1;102(4):e208048.

54. 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. Available from: http://publications.aap.org/pediatrics/article-pdf/142/6/e20180931/1075531/peds_20180931.pdf

55. Wikipedia the free encyclopedia. Document-term matrix. [Internet]. Available from: https://en.wikipedia.org/wiki/Document-term_matrix. [Accessed 28th July 2024].

56. Wikipedia the free encyclopedia. Bag-of-words model. [Internet]. Available from: https://en.wikipedia.org/wiki/Bag-of-words_model. [Accessed 28th July 2024].

57. IBM. What are Naïve Bayes classifiers? [Internet]. Available from: https://www.ibm.com/topics/naive-bayes#:~:text=The%20Na%C3%AFve%20Bayes%20classifier%20is,probability%20to%20perform%20classification%20tasks. [Accessed 18th July 2024].

58. Wikipedia the free encyclopedia. F-score. [Internet]. Available from: https://en.wikipedia.org/wiki/F-score. [Accessed 28th July 2024].

59. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

60. Simplilearn. Top Deep Learning Interview Questions and Answers for 2024. Available from: https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-interview-questions. [Accessed 9th August 2024].

61. 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. 2021 Oct 1;91:141–5.

62. Learndatasci.com. TF-IDF: Term Frequency – Inverse Document Frequency. [Internet]. Available from: https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/#:~:text=Term%20Frequency%20%2D%20Inverse%20Document%20Frequency%20(TF%2DIDF)%20is,%2C%20relative%20to%20a%20corpus. [Accessed 9th August 2024].

63. Simplilearn. What is a Chi-Square Test? Formula, Examples & Application. [Internet]. Available from: https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test. [Accessed 9th August 2024].

Appendix

Appendix 1: removed stop words

Table A1. List of removed stop words.

Removed stop words
'de', 'het', 'een', 'van', 'naar', 'op', 'in', 'voor', 'met', 'uit', 'over', 'tot', 'door', 'aan', 'na', 'per', 'sinds', 'bij', 'boven', 'onder', 'om', 'via', 'ons', 'u', 'we', 'mij', 'me', 'zij', 'ze', 'hij', 'hem', 'haar', 'jullie', 'je', 'jezelf', 'jouw', 'hun', 'hen', 'wij', 'onze', 'zich', 'iemand', 'niemand', 'patient', 'patiënt', 'pt', 'zichzelf', 'zijn', 'ben', 'bent', 'is', 'was', 'waren', 'wezen', 'worden', 'word', 'wordt', 'werd', 'werden', 'geworden', 'doen', 'doet', 'deden', 'deed', 'gedaan', 'maken', 'maakt', 'maakte', 'maakten', 'gemaakt', 'zullen', 'zal', 'zult', 'zou', 'zouden', 'moeten', 'moet', 'gemoeten', 'moest', 'willen', 'wil', 'wilt', 'wilde', 'wilden', 'wou', 'wouden', 'mogen', 'mag', 'mocht', 'mochten', 'hebben', 'heb', 'hebt', 'heeft', 'had', 'hadden', 'gehad', 'kan', 'kunt', 'kunnen', 'kon', 'konden', 'gekund', 'gaan', 'ga', 'gaat', 'ging', 'gingen', 'gegaan', 'kom', 'komt', 'komen', 'kwam', 'kwamen', 'gekomen', 'en', 'dat', 'die', 'maar', 'als', 'dan', 'ook', 'of', 'dus', 'omdat', 'indien', 'echter', 'alsook', 'evenals', 'eveneens', 'mede', 'bovendien', 'terwijl', 'er', 'hier', 'daar', 'te', 'nog', 'dus', 'hoe', 'welke', 'hierdoor', 'hiermee', 'waardoor', 'waarvoor', 'waarom', 'waarbij', 'daarom', 'vervolgens', 'daarnaast', 'daarna', 'tevens', 'eens', ' reeds', 'toch', 'al', 'opnieuw', 'nogmaals', 'meer', 'minder', 'veel', 'weinig', 'zeer', 'vaak', 'soms', 'zelden', 'altijd', 'nooit', 'alleen', 'hierna', 'ooit', 'toen', 'nimmer', 'januari', 'februari', 'maart', 'april', 'mei', 'juni', 'juli', 'augustus', 'september', 'oktober', 'november', 'december', 'maandag', 'dinsdag', 'woensdag', 'donderdag', 'vrijdag', 'zaterdag', 'zondag', 'reden', 'reden_van_komst', 'reden_van_verwijzing', 'anamnese', 'lichamelijk_onderzoek', 'beleid', 'conclusie', 'andere', 'anderen', 'ander', 'anders', 'deze', 'dit', 'iets', 'niets', 'wie', 'wat', 'men', 'alles', 'alle', 'sommige', 'sommigen', 'mg', 'cg', 'dg', 'g', 'kg', 'ml', 'cl', 'dl', 'l', 'mm', 'cm', 'dm', 'm', 'km', 'ja', 'nee', 'niet', 'wel', 'geen', 'want', 'hoewel', 'mits', 'tenzij', 'aangezien', 'voordat', 'nadat', 'tijdens', 'zodra', 'ofschoon', 'alhoewel', 'nu', 'op_de', 'in_de', 'aan_de', 'naar_de', 'over_de', 'op_het', 'in_het', 'aan_het', 'naar_het', 'over_het', 'op_een', 'in_een', 'aan_een', 'naar_een', 'over_een', 'van_de', 'van_het', 'van_een', 'met_de', 'met_het', 'met_een', 'voor_de', 'voor_het', 'voor_een', 'na_de', 'na_het', 'na_een', 'bij_een', 'bij_de', 'bij_het', 'is_er', 'er_is'

Appendix 2: baseline table

Table A2. 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.

Appendix 3: analyses of the most important features

Word-frequency analysis

Word-frequency analysis was conducted to identify the most common features within the dataset. The top 50 most frequently occurring features are presented in Figure A1.

	feature	frequency	rank	docfreq	group
1	moeder	4992	1	1339	all
2	symmetrisch	3896	2	1316	all
3	goed	3267	3	1291	all
4	normaal	2606	4	1208	all
5	ouders	2367	5	1058	all
6	epilepsie	2340	6	1394	all
7	onderzoek	2328	7	1538	all
8	jaar	2309	8	1062	all
9	vader	2309	8	1024	all
10	aanval	2223	10	839	all
11	aanvallen	2161	11	848	all
12	normale	1940	12	959	all
13	armen	1931	13	1107	all
14	weer	1902	14	1003	all
15	ontwikkeling	1878	15	1142	all
16	benen	1841	16	1085	all
17	1	1806	17	978	all
18	ogen	1802	18	1002	all
19	2	1600	19	919	all
20	voorgeschiedenis	1600	19	1455	all
21	intact	1595	21	824	all
22	medicatie	1531	22	1279	all
23	minuten	1490	23	822	all
24	reden_van	1439	24	1433	all
25	maanden	1349	25	776	all
26	contact	1328	26	839	all
27	dag	1301	27	816	all
28	5	1285	28	686	all
29	beiderzijds	1265	29	872	all
30	reflexen	1248	30	1194	all
31	keer	1247	31	714	all
32	school	1247	31	664	all
33	gelaat	1228	33	1050	all
34	armen_en	1203	34	857	all
35	en_benen	1138	35	826	all
36	kracht	1124	36	956	all
37	ongestoord	1123	37	652	all
38	3	1098	38	713	all
39	zwangerschap	1089	39	804	all
40	weken	1055	40	737	all
41	familie	1055	40	817	all
42	hoofd	1049	42	633	all
43	motoriek	1025	43	776	all
44	lichamelijk	1020	44	1003	all
45	pupillen	1020	44	988	all
46	hersenzenuwen	999	46	980	all
47	koorts	992	47	531	all
48	sensibiliteit	991	48	733	all
49	goede	987	49	682	all
50	lopen	981	50	676	all

Figure A1. The top 50 most frequently occurring features in the dataset. Feature = the word or term identified in the dataset, frequency = the total number of occurrences of the feature across all documents, rank = the rank of the feature based on its frequency, Docfreq = the number of documents in which the feature appears, group = the category or class to which the feature belongs (in this case ‘epilepsy’ and ‘no epilepsy). "Moeder" is the most frequently occurring feature in the dataset, appearing 4,992 times.

TF-IDF frequency analysis

TI-IDF frequency analysis of the training set identified several key features based on their TF-IDF scores across the letters. TF-IDF is a statistical measure that indicates the importance of a term in a letter relative to the collection of letters.[62] It combines term frequency (TF), how often a term appears in a specific letter, and inverse document frequency (IDF), which is how rare or common the feature is across the letters.[62] In Analysis 1, the most significant features are “aanval ”, “school”, “schokken”, “links”, and “ongeveer”, each demonstrating high frequencies and indicating their importance in the dataset (as illustrated in Figure A1). For instance, the feature “aanval“ has the highest frequency of 416.8640 and appears in 600 letters, underscoring its significance. This high TF-IDF score suggests that “aanval” is a key feature in the letters where it frequently appears, although it is not present in all letters, contributing to its high score. The top 30 most frequently occurring features are presented in Figure A2.

	feature	frequency	rank	docfreq	group
1	aanval	416.8640	1	600	all
2	school	327.1643	2	457	all
3	schokken	291.1789	3	371	all
4	links	282.7838	4	312	all
5	ongeveer	249.9295	5	435	all

Figure A2. Frequency analysis on the TF-IDF DFM of the training set. The word-frequency analysis applying TF-IDF to the DFM demonstrates the most important features based on their adjusted frequency across letters. Feature = the word or term being analysed, frequency = the TF-IDF adjusted frequency of the feature, rank = the rank of the feature based on its frequency, docfreq = the number of letters in which the feature appears, group = the group to which the feature belongs (“all” indicating the entire training set).

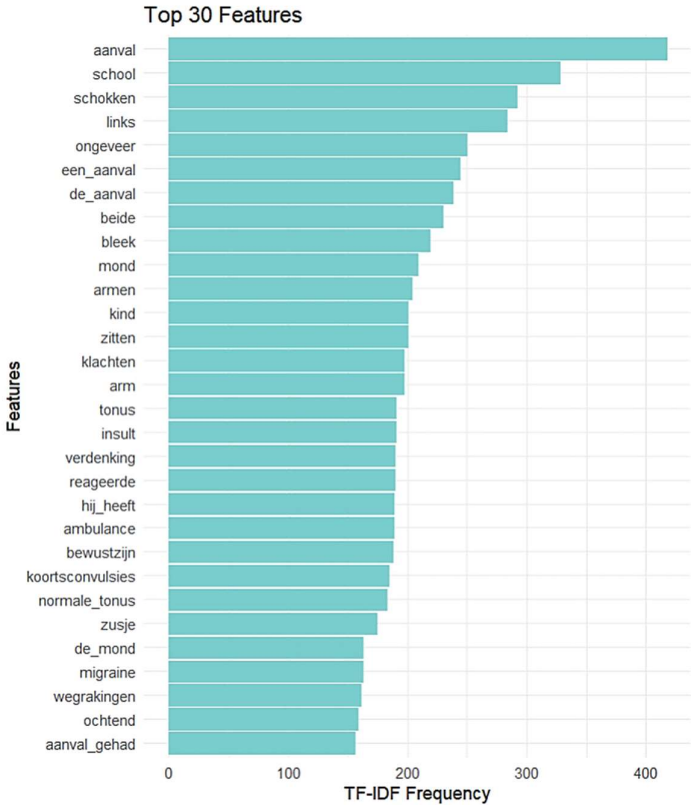

Figure A3. The top 30 most frequently occurring features. The figure presents the top 30 features from the corpus ranked based on their TF-IFD frequency. TF-IDF is a statistical measure that quantifies the importance of a feature in a letter relative to a collection of letters.[62] The higher the TF-IDF frequency, the more significant the feature is within the context of the letters. “aanval” is the most frequent feature with the highest TF-IDF frequency. “school” is the second most frequent feature. “schokken”, “links”, “ongeveer” are other significant features, ranked in descending order of TF-IDF frequency. Each bar represents a feature and the length of the bar corresponds to the TF-IDF frequency of that feature.

In Analysis 2, the most significant features are "aanval ", "school", "schokken", "links", and "een_aanval", each demonstrating high frequencies and indicating their importance in the dataset (as illustrated in Figure A4). For instance, the feature "aanval" has the highest TF-IDF score of 489.7944 and appears in 679 letters, underscoring its relevance. This high TF-IDF score suggests that "aanval" is a key feature in the letters where it frequently appears, although it is not present in all letters, contributing to its high score. The top 30 most frequently occurring features are presented in Figure A5.

	feature	frequency	rank	docfreq	group
1	aanval	489.7944	1	679	all
2	school	387.2997	2	525	all
3	schokken	341.3493	3	434	all
4	links	335.0784	4	373	all
5	een_aanval	293.3705	5	306	all

Figure A4. Frequency analysis on the TF-IDF DFM of the training set. The word-frequency analysis applying TF-IDF to the DFM demonstrates the most important features based on their adjusted frequency across the letters. Feature = the word or term being analysed, frequency = the TF-IDF adjusted frequency of the feature, rank = the rank of the feature based on its frequency, docfreq = the number of letters in which the feature appears, group = the group to which the feature belongs ("all" indicating the entire training set).

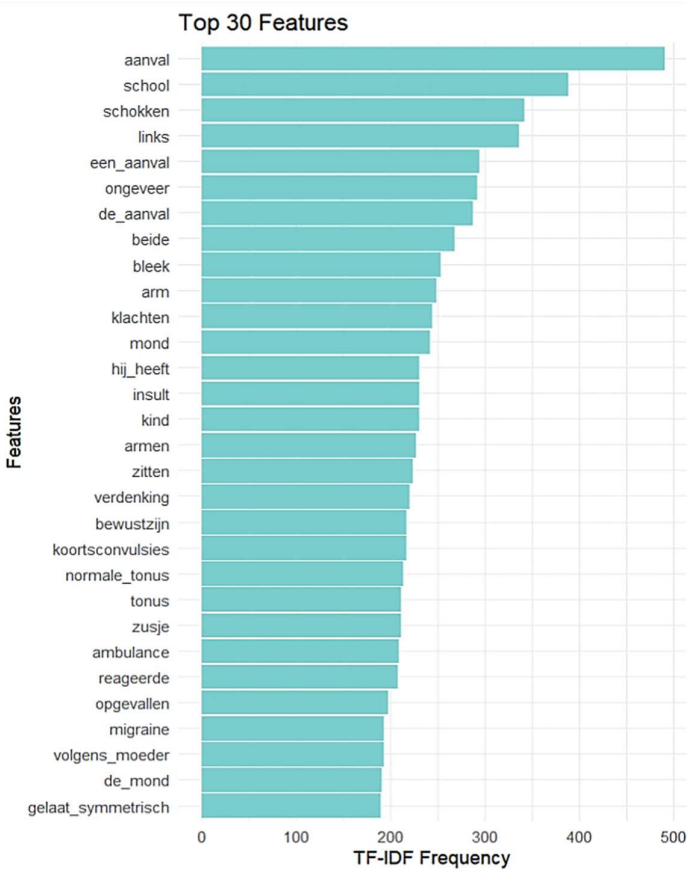

Figure A5. The top 30 most frequently occurring features. The figure presents the top 30 features from the text corpus ranked based on their TF-IFD frequency. TF-IDF is a statistical measure that quantifies the importance of a feature in a letter relative to a collection of letters.[62] The higher the TF-IDF frequency, the more significant the feature is within the context of the letters. "aanval" is the most frequent feature with the highest TF-IDF frequency. "school" is the second most frequent feature. "schokken", "links", "een_aanval" are other significant features, ranked in descending order of TF-IDF frequency. Each bar represents a feature and the length of the bar corresponds to the TF-IDF frequency of that feature.

Comparative word-frequency analysis

Comparative word-frequency analysis identified several features that significantly differ in frequency between the 'yes' group (epilepsy) and the 'no' group (no epilepsy). The Chi-squared statistic measures how much the observed frequency of the feature differs from the expected frequency under the null hypothesis of no difference between groups.[63] Higher Chi-squared values indicate a greater difference from what would be expected if there was no association between the feature and the groups.[63] A high Chi-squared value does not necessarily mean that the difference in feature frequency between groups is large, it only indicates that the observed frequencies are significantly different from the expected frequencies. Features with the highest Chi-squared value in Analysis 1 included "aanval", "insult", "midazolam", "de_aanval", and "een_aanval". The high Chi-squared values indicate their importance in differentiating the two groups (as illustrated in Figure A6). For instance, the feature "aanval" appears 766 times in the 'yes' group (target group) compared to 832 times in the 'no' group (reference group), with a Chi-squared value of 161.88267, demonstrating a significant disparity.

	feature	chi2	p	n_target	n_reference
1	aanval	161.88267	0.000000e+00	766	832
2	insult	89.75152	0.000000e+00	151	95
3	midazolam	46.44194	9.437340e-12	78	49
4	de_aanval	46.22549	1.053968e-11	174	173
5	een_aanval	40.35811	2.114254e-10	186	200
6	en_ogen	37.24490	1.041869e-09	33	10
7	mond	36.76283	1.334094e-09	159	167
8	neusspray	36.57207	1.471248e-09	60	37
9	de_mond	34.45516	4.361842e-09	105	95
10	links	34.39310	4.503163e-09	234	285
11	linkerarm	34.26719	4.804142e-09	37	15
12	ambulance	32.99029	9.262042e-09	121	119
13	de_ochtend	32.26153	1.347551e-08	62	43
14	naar_links	30.66384	3.068302e-08	69	53
15	arm	30.63362	3.116466e-08	112	110
16	midazolam_neusspray	30.13581	4.028247e-08	49	30
17	schokken	28.50387	9.351124e-08	267	353
18	ochtend	27.79505	1.348709e-07	96	92
19	epileptisch_insult	26.23879	3.017016e-07	31	14
20	eerste_insult	25.89233	3.610008e-07	23	7
21	aanval_gehad	25.39811	4.663747e-07	92	90
22	hij_had	25.00740	5.711082e-07	74	66
23	geen_koorts	24.74065	6.558635e-07	92	91
24	ogen_naar	24.59924	7.057953e-07	66	56
25	bed	24.24887	8.465692e-07	191	242
26	hoorde	23.75625	1.093397e-06	59	48
27	slijm	23.57582	1.200870e-06	24	9
28	geslapen	23.48410	1.259502e-06	68	60
29	was_hij	23.28531	1.396624e-06	105	112
30	aanvallen	22.69336	1.900230e-06	589	932

Figure A6. Comparative word-frequency analysis of the training set. The comparative word-frequency analysis identifies the features that significantly differ in frequency between the target group ('yes'-group or 'epilepsy') and the reference group ('no'-group or 'no epilepsy'). Feature = the word or term being analysed, n_target = the frequency of the feature in the target group, n_reference = the frequency of the feature in the reference group.

In Analysis 2, features with the highest Chi-squared value included "aanval", "insult", "de_aanval", "midazolam", and "mond". The high Chi-squared values indicate their importance in differentiating the two groups (as illustrated in Figure A7). For instance, the feature "aanval" appears 921 times in the 'yes' group (target group) compared to 927 times in the 'no' group (reference group), with a Chi-squared value of 224.09850, demonstrating a significant disparity.

	feature	chi2	p	n_target	n_reference
1	aanval	224.09850	0.000000e+00	921	927
2	insult	131.44244	0.000000e+00	191	103
3	de_aanval	93.77799	0.000000e+00	234	186
4	midazolam	56.37867	5.973000e-14	94	57
5	mond	49.29583	2.201350e-12	192	189
6	arm	45.34385	1.653067e-11	157	147
7	ambulance	44.16363	3.020406e-11	134	118
8	schokken	43.34812	4.581757e-11	333	410
9	neusspray	42.33639	7.684919e-11	68	40
10	een_aanval	41.47442	1.194236e-10	227	253
11	midazolam_neusspray	40.45234	2.014714e-10	61	34
12	seh	38.28773	6.104528e-10	98	79
13	links	37.27801	1.024333e-09	286	352
14	de_mond	37.05344	1.149355e-09	123	113
15	rechts	34.65007	3.946247e-09	290	365
16	hoorde	32.90546	9.675136e-09	67	47
17	eerste_insult	32.88469	9.779036e-09	26	6
18	komst_eerste	30.71062	2.995210e-08	19	2
19	kwijlen	30.05103	4.208248e-08	35	15
20	insult_anamnese	30.02039	4.275275e-08	46	26
21	geslapen	29.95960	4.411432e-08	72	56
22	epileptisch_insult	28.40971	9.817150e-08	42	23
23	de_ambulance	28.06369	1.173880e-07	79	67
24	speeksel	27.95708	1.240363e-07	27	9
25	minuten	27.85337	1.308662e-07	432	619
26	ochtend	27.77156	1.365182e-07	115	116
27	hij_had	27.60417	1.488575e-07	90	82
28	de_avond	27.14875	1.883871e-07	43	25
29	linkerarm	26.83734	2.213190e-07	38	20
30	rechter_arm	23.64156	1.160538e-06	44	29

Figure A7. Comparative word-frequency analysis of the training set. The comparative word-frequency analysis identifies the features that significantly differ in frequency between the target group ('yes'-group or 'epilepsy') and the reference group ('no'-group or 'no epilepsy'). Feature = the word or term being analysed, n_target = the frequency of the feature in the target group, n_reference = the frequency of the feature in the reference group.

Key features for classification

Analysis 1

The top 30 features considered most important by the model for classification are demonstrated in Figures A8 and A9. Features contributing to the classification of a diagnosis of 'epilepsy' included "en_ogen", "eerste_insult", "speeksel", "komt_eerst" and, "112_gebeld" (Figure A8). Features contributing to the classification of a diagnosis of 'no epilepsy' included "wegrakingen", "half_uur", "flauwvallen", "oorzaak", and "klachten" (Figure A9).

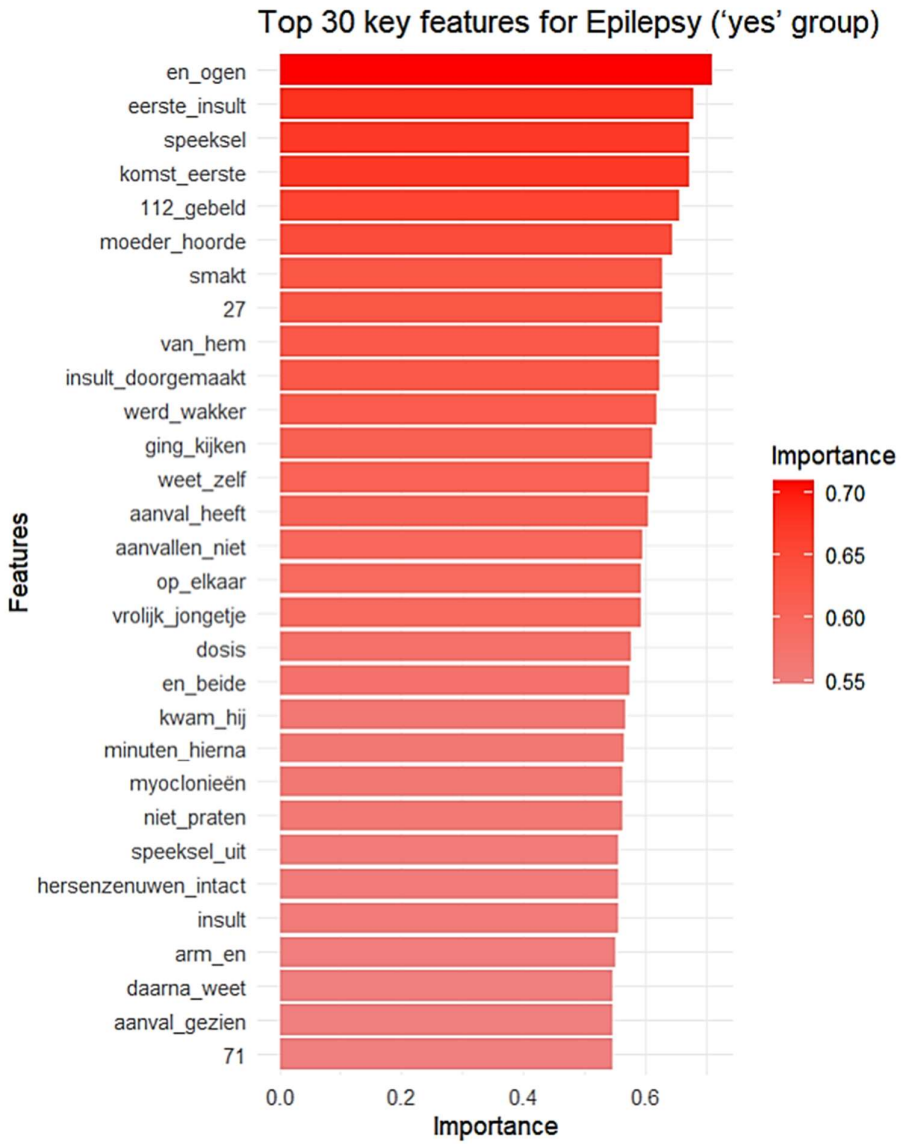

Figure A8. The top 30 key features for the classification of epilepsy. The figure presents the top 30 features of epilepsy, ranked by their relative importance scores. The importance score quantifies the relevance of each feature in distinguishing epilepsy from no epilepsy. "en_ogen" is the most important feature, indicating its strong association with epilepsy. "eerste_insult", "speeksel" are other significant features, listed in descending order of importance. Each bar represents a feature and the length of the bar corresponds to its importance score, demonstrating how important each feature is in the context of epilepsy. The colour gradient further highlights the importance, with darker shades indicating higher importance.

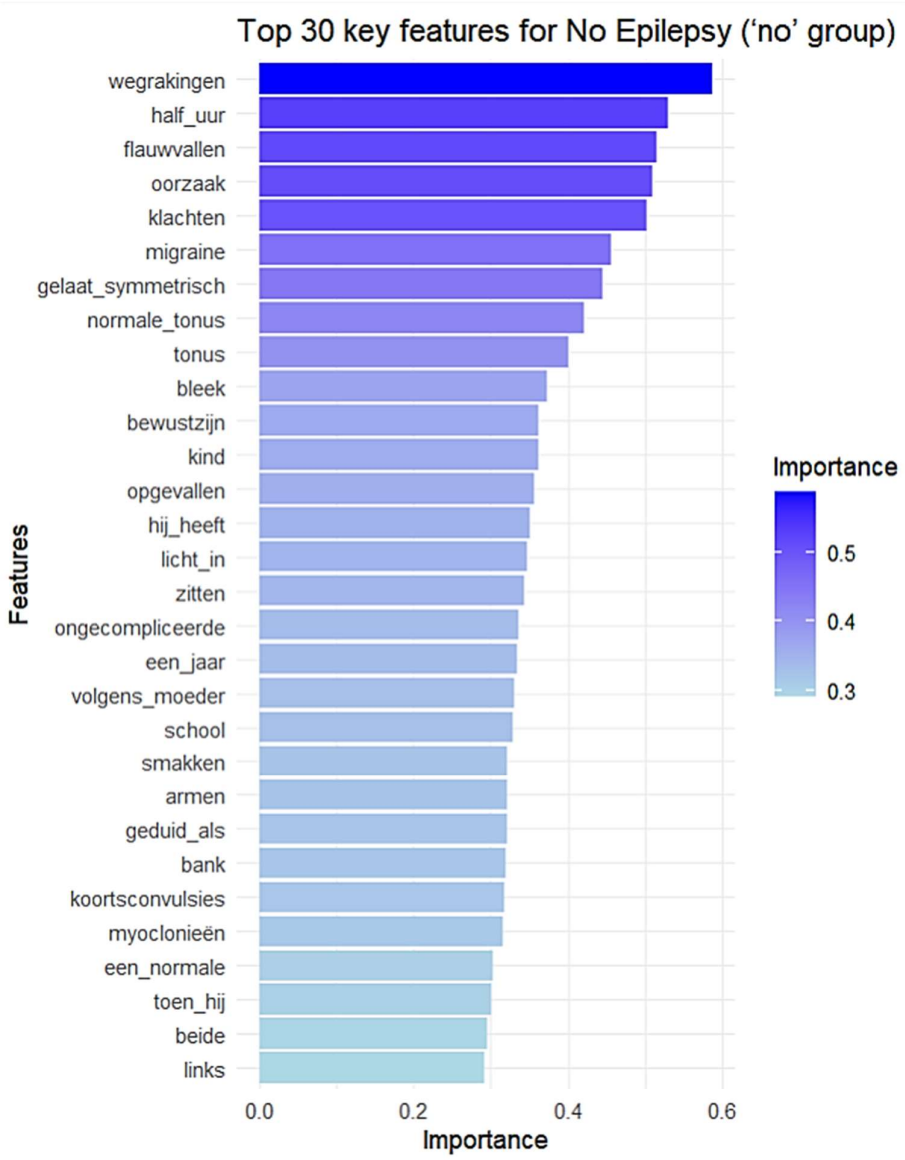

Figure A9. The top 30 key features for the classification of 'no epilepsy'. The figure presents the top 30 features of no epilepsy, ranked by their relative importance scores. The importance score quantifies the relevance of each feature in distinguishing no epilepsy from epilepsy. "wegrakingen" is the most important feature, indicating its strong association with no epilepsy. "half_uur", "flauwvallen" are other significant features, listed in descending order of importance. Each bar represents a feature and the length of the bar corresponds to its importance score, demonstrating how important each feature is in the context of no epilepsy. The colour gradient further highlights the importance, with darker shades indicating higher importance.

Analysis 2

The top 30 features considered most important by the model for classification are demonstrated in Figures A10 and A11. Features contributing to the classification of a diagnosis of 'epilepsy' included "komst_eerste", "eerste_insult", "speeksel", "werd_wakker" and, "112_gebeld" (Figure A10). Features contributing to the classification of a diagnosis of 'no epilepsy' included "flauwvallen", "klachten", "wegrakingen", "oorzaak", and "migraine" (Figure A11).

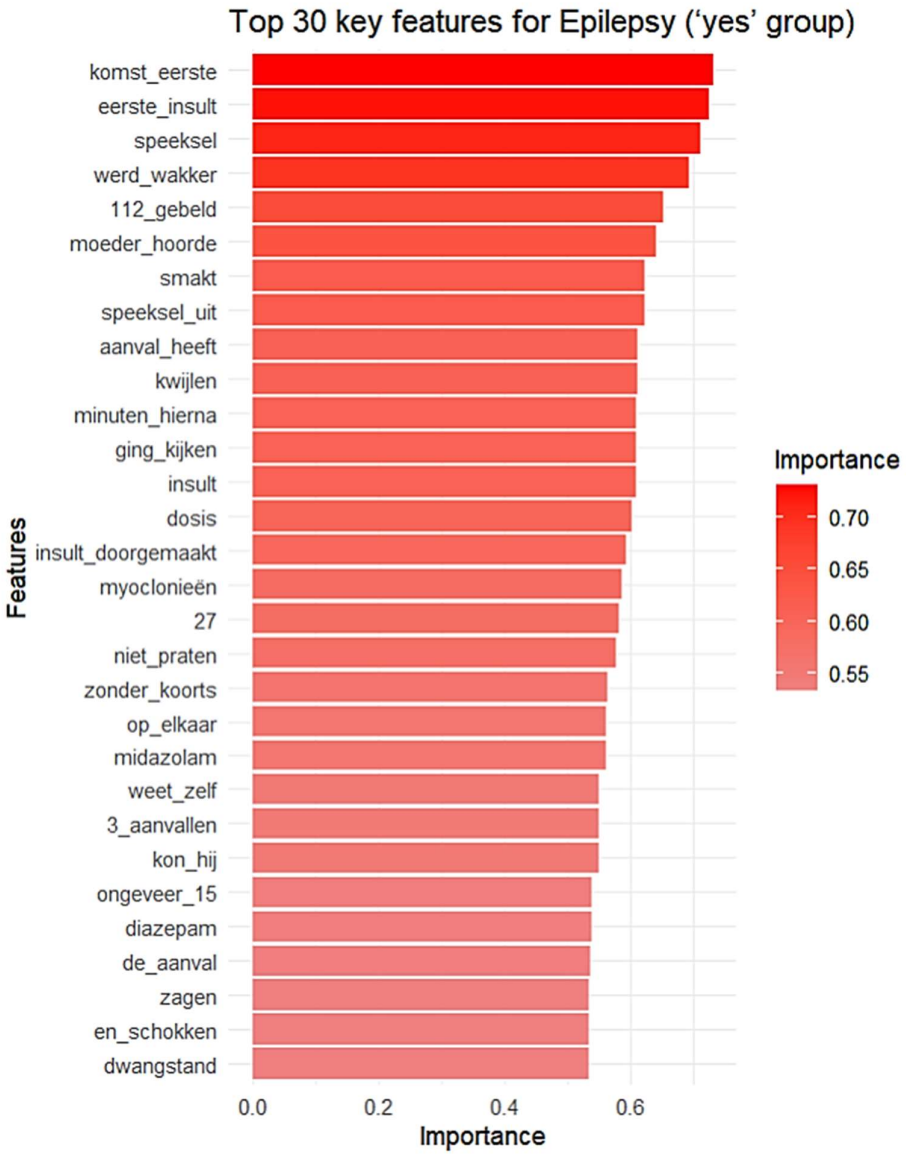

Figure A10. The top 30 key features for the classification of epilepsy. The figure presents the top 30 features of epilepsy, ranked by their relative importance scores. The importance score quantifies the relevance of each feature in distinguishing epilepsy from no epilepsy. "komst_eerste" is the most important feature, indicating its strong association with epilepsy. "eerste_insult", "speeksel" are other significant features, listed in descending order of importance. Each bar represents a feature and the length of the bar corresponds to its importance score, demonstrating how important each feature is in the context of epilepsy. The colour gradient further highlights the importance, with darker shades indicating higher importance.

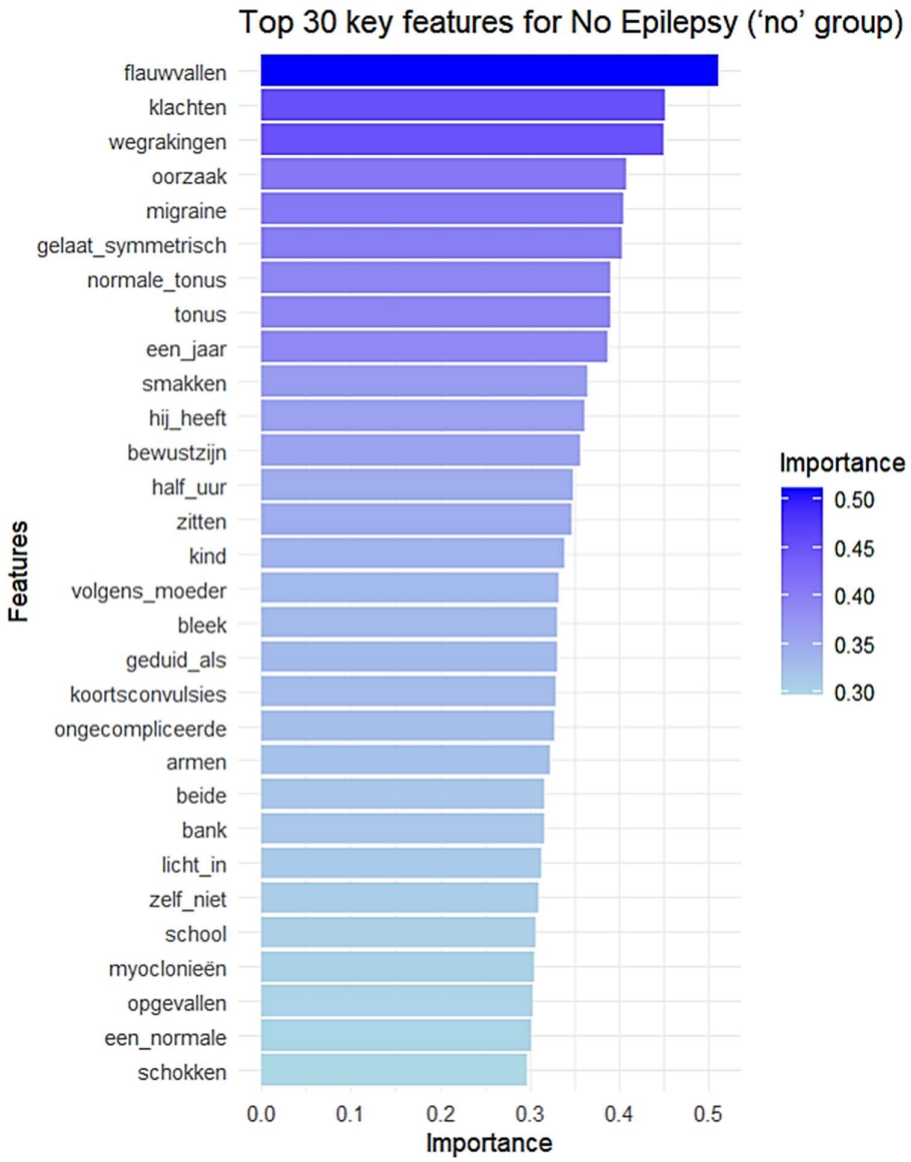

Figure A11. The top 30 key features for the classification of 'no epilepsy'. The figure presents the top 30 features of no epilepsy, ranked by their relative importance scores. The importance score quantifies the relevance of each feature in distinguishing no epilepsy from epilepsy. "flauwvallen" is the most important feature, indicating its strong association with no epilepsy. "klachten", "wegrakingen" are other significant features, listed in descending order of importance. Each bar represents a feature and the length of the bar corresponds to its importance score, demonstrating how important each feature is in the context of no epilepsy. The colour gradient further highlights the importance, with darker shades indicating higher importance.

Appendix 4: confusion matrices and graphical representations of results

Analysis 1A

	Predicted class		
Actual class	No	Unclear	Yes
No	566	12	93
Unclear	16	41	6
Yes	99	31	230

	Predicted class		
Actual class	No	Unclear	Yes
No	223	21	43
Unclear	17	2	7
Yes	61	12	81

Figure A12. The confusion matrix of the training set (left) and the test set (right). The confusion matrices represent the classification model's performance based on actual and predicted classes. The rows represent the actual classes and the columns represent the predicted classes. For the training set, the model correctly classified 566 cases as 'no' and 230 cases as 'yes'. Erroneously, 93 cases that were actually 'no' were classified as 'yes', and 99 cases that were actually 'yes' were classified as 'no'. Additionally, 16 cases that were actually 'unclear' were classified as 'no', 41 cases that were actually 'unclear' were classified as 'unclear', and 6 cases that were actually 'unclear' were classified as 'yes'. The positive predictive value for the 'yes' group is 0.64 and the negative predictive value is 0.87. The positive predictive value for the 'no' group is 0.84 and the negative predictive value is 0.73. For the test set, the model correctly classified 223 cases as 'no' and 81 cases as 'yes'. Erroneously, 43 cases that were actually 'no' were classified as 'yes', and 61 cases that were actually 'yes' were classified as 'no'. Additionally, 17 cases that were actually 'unclear' were classified as 'no', 2 cases that were actually 'unclear' were classified as 'unclear', and 7 cases that were actually 'unclear' were classified as 'yes'. The positive predictive value for the 'yes' group is 0.53 and the negative predictive value is 0.84. The positive predictive value for the 'no' group is 0.78 and the negative predictive value is 0.57.

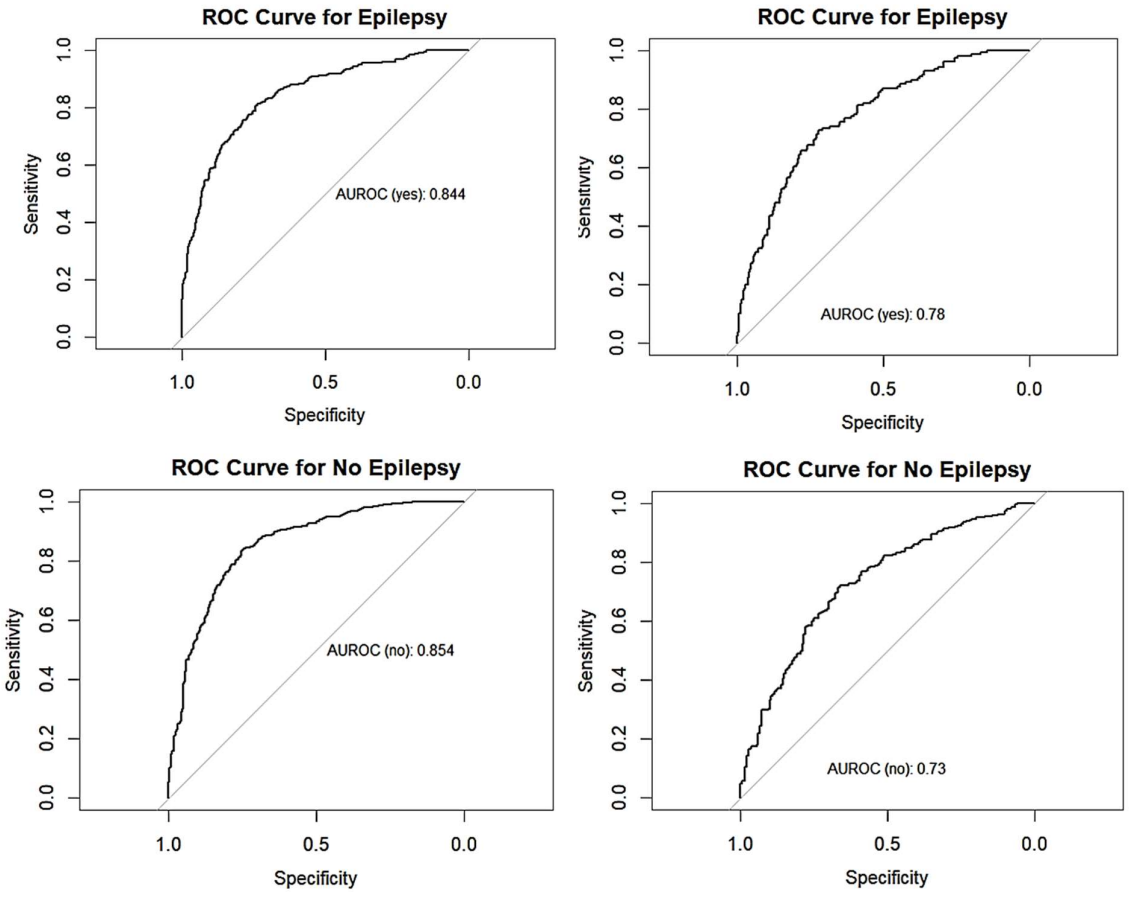

Figure A13. ROC curves of the classification model. The top row presents the performance for predicting 'epilepsy' in the training set (left) and the test set (right). The bottom row presents the performance for predicting 'no epilepsy' in the training set (left) and test set (right). The AUROC values represent the model's predictive accuracy, with higher values indicating better performance.

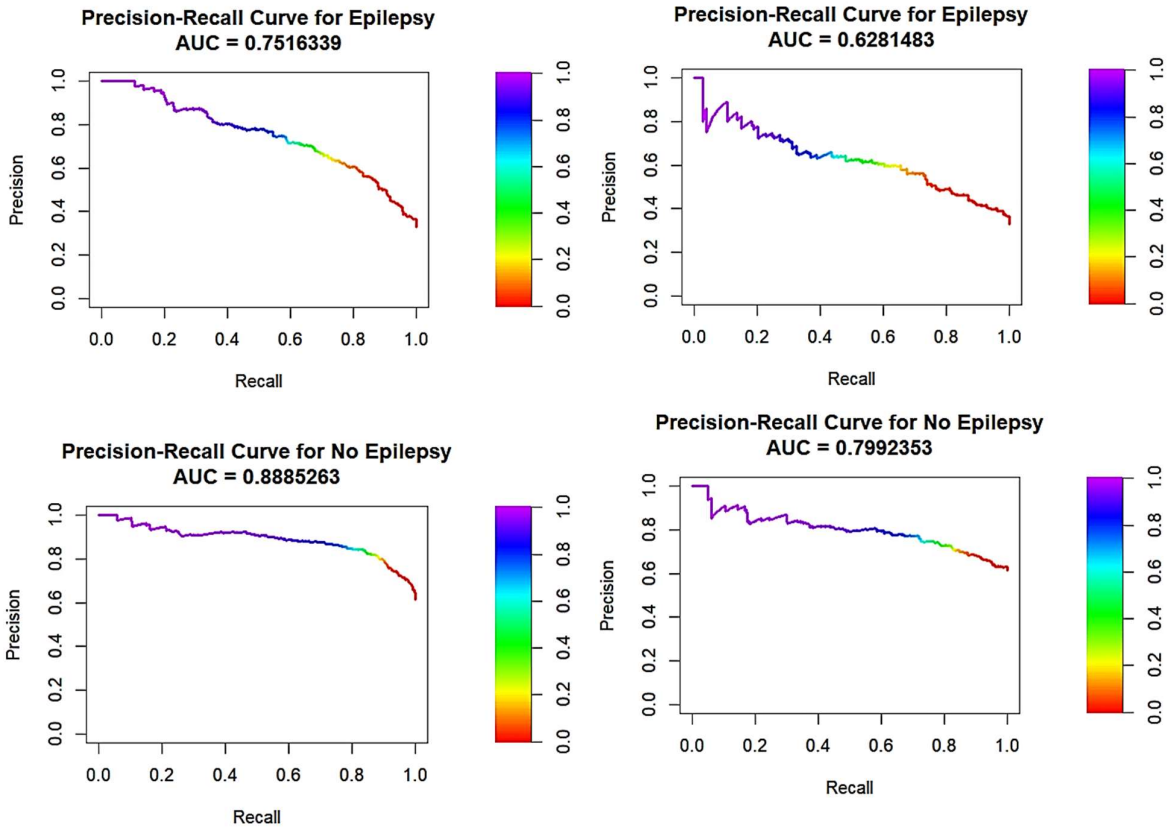

Figure A14. Precision-Recall Curves of the classification model. The top row demonstrates the performance for predicting 'epilepsy' in the training set (left) and the test set (right). The bottom row demonstrates the performance for predicting 'no epilepsy' in the training set (left) and the test set (right). The AUPRC values represent the trade-off between precision and recall, with higher values indicating better performance. The colour gradient represents different threshold values.

Analysis 1B

	Predicted class		
Actual class	No	Unclear	Yes
No	173	7	39
Unclear	32	41	11
Yes	34	14	49

Figure A15. The confusion matrix of the training set (left) and the test set (right). The confusion matrices represent the classification model's performance based on actual and predicted classes. The rows represent the actual classes and the columns represent the predicted classes. For the test set, the model correctly classified 173 unclear cases as 'no' and 49 unclear cases as 'yes'. Erroneously, 39 unclear cases that were finally 'no' were classified as 'yes', and 34 unclear cases that were finally 'yes' were classified as 'no'. Additionally, 32 cases that were finally still 'unclear' were classified as 'no', 41 cases that were finally still 'unclear' were classified as 'unclear', and 11 cases that were finally still 'unclear' were classified as 'yes'. The positive predictive value for the 'yes' group is 0.51 and the negative predictive value is 0.84. The positive predictive value for the 'no' group is 0.80 and the negative predictive value is 0.64.

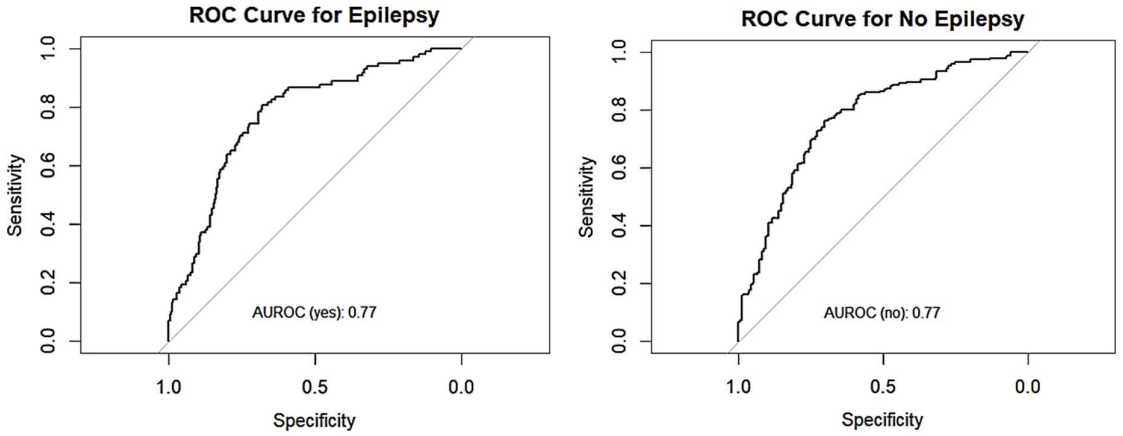

Figure A16. ROC curves of the classification model. The curves present the performance for predicting 'epilepsy' and 'no epilepsy' in the test set. The AUROC values represent the model's predictive accuracy, with higher values indicating better performance.

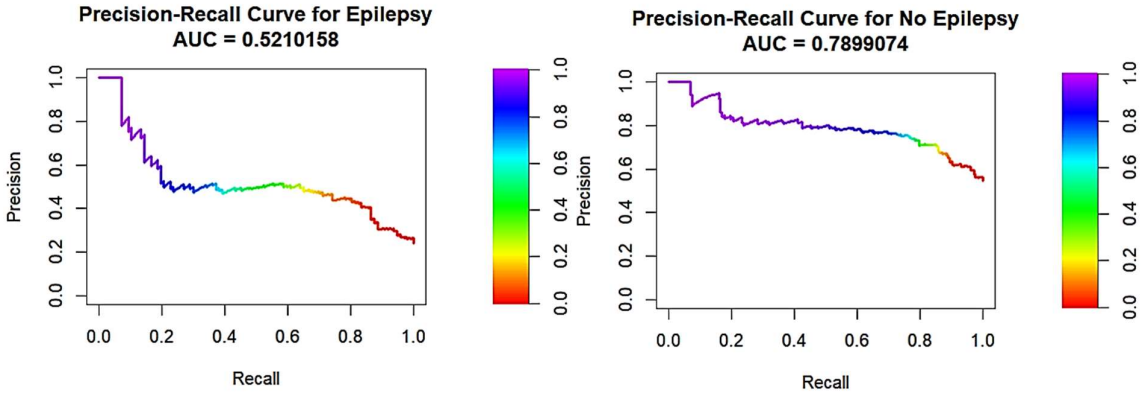

Figure A17. Precision-Recall Curves of the classification model. The curves present the performance for predicting 'epilepsy' and 'no epilepsy' in the test set. The AUPRC values represent the trade-off between precision and recall, with higher values indicating better performance. The colour gradient represents different threshold values.

Analysis 2

	Predicted class		
Actual class	No	Unclear	Yes
No	622	33	112
Unclear	23	39	8
Yes	116	31	266

	Predicted class		
Actual class	No	Unclear	Yes
No	161	13	17
Unclear	8	8	3
Yes	43	15	43

Figure A18. The confusion matrix of the training set (left) and the test set (right). The confusion matrices represent the classification model's performance based on actual and predicted classes. The rows represent the actual classes and the columns represent the predicted classes. For the training set, the model correctly classified 622 cases as 'no' and 266 cases as 'yes'. Erroneously, 112 cases that were actually 'no' were classified as 'yes', and 116 cases that were actually 'yes' were classified as 'no'. Additionally, 23 cases that were actually 'unclear' were classified as 'no', 39 cases that were actually 'unclear' were classified as 'unclear', and 8 cases that were actually 'unclear' were classified as 'yes'. The positive predictive value for the 'yes' group is 0.64 and the negative predictive value is 0.86. The positive predictive value for the 'no' group is 0.81 and the negative predictive value is 0.71. For the test set, the model correctly classified 161 cases as 'no' and 43 cases as 'yes'. Erroneously, 17 cases that were actually 'no' were classified as 'yes', and 43 cases that were actually 'yes' were classified as 'no'. Additionally, 8 cases that were actually 'unclear' were classified as 'no', 8 cases that were actually 'unclear' were classified as 'unclear', and 3 cases that were actually 'unclear' were classified as 'yes'. The positive predictive value for the 'yes' group is 0.43 and the negative predictive value is 0.90. The positive predictive value for the 'no' group is 0.84 and the negative predictive value is 0.58.

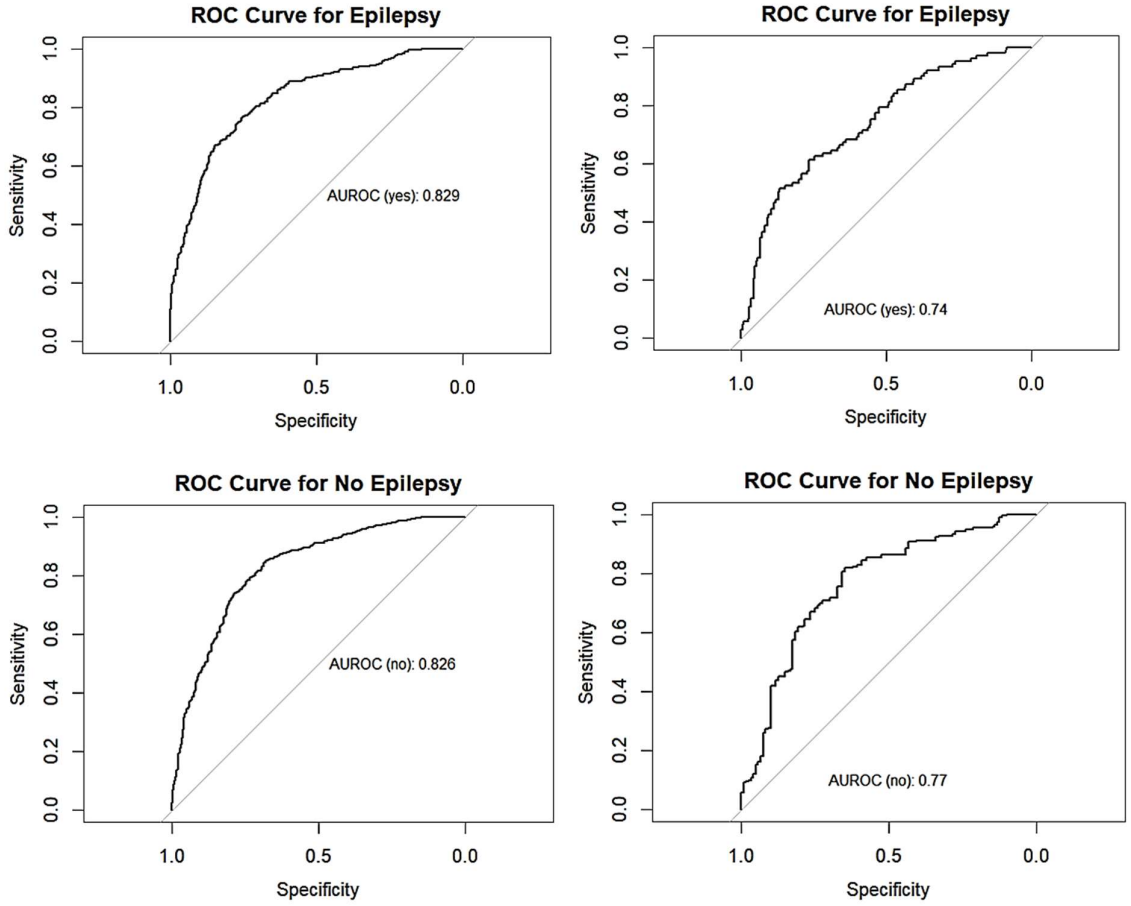

Figure A19. ROC curves of the classification model. The top row presents the performance for predicting 'epilepsy' in the training set (left) and the test set (right). The bottom row presents the performance for predicting 'no epilepsy' in the training set (left) and test set (right). The AUROC values represent the model's predictive accuracy, with higher values indicating better performance.

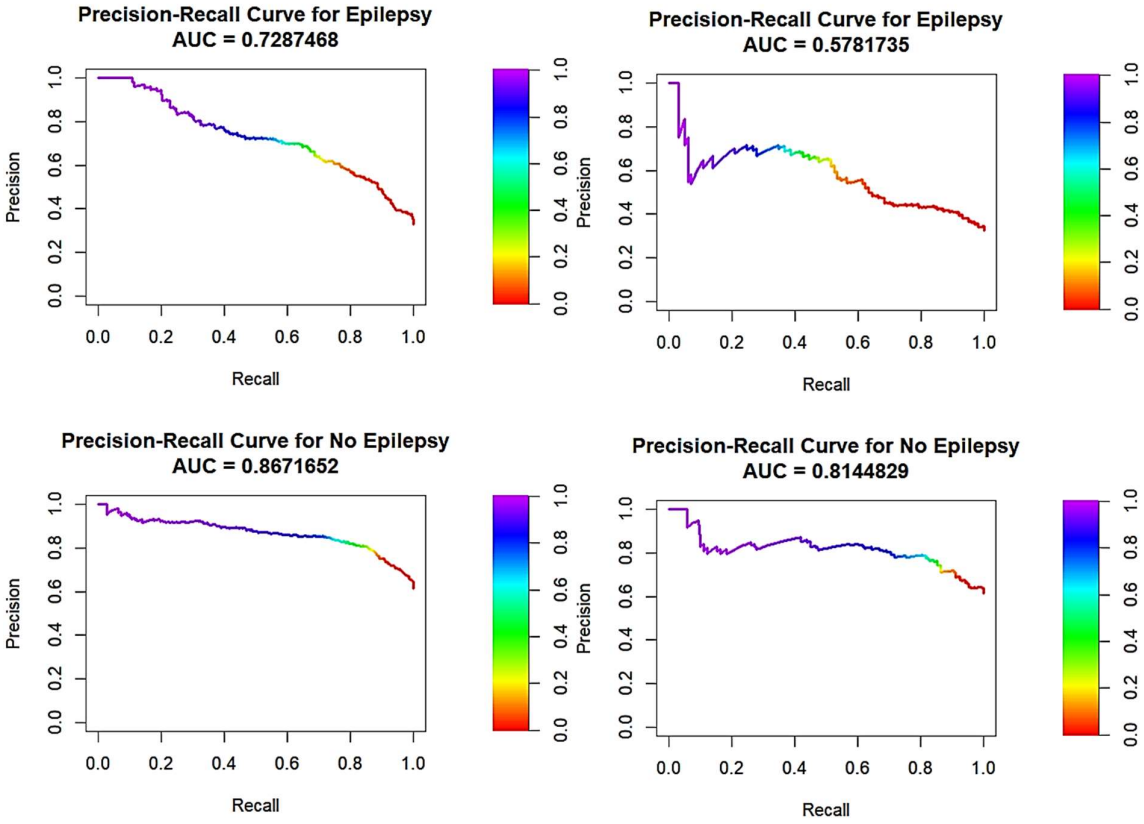

Figure A20. Precision-Recall Curves of the classification model. The top row demonstrates the performance for predicting 'epilepsy' in the training set (left) and the test set (right). The bottom row demonstrates the performance for predicting 'no epilepsy' in the training set (left) and the test set (right). The AUPRC values represent the trade-off between precision and recall, with higher values indicating better performance. The colour gradient represents different threshold values.