Seizure: European Journal of Epilepsy 114 (2024) 84–89

Contents lists available at ScienceDirect

Seizure: European Journal of Epilepsy

journal homepage: www.elsevier.com/locate/seizure

Predicting the cause of seizures using features extracted from interactions with a virtual agent

Nathan Pevy[a,*], Heidi Christensen[b], Traci Walker[c], Markus Reuber[d]

[a] Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
[b] Department of Computer Science, University of Sheffield, Sheffield, UK
[c] Division of Human Communication Sciences, University of Sheffield, Sheffield, UK
[d] Academic Neurology Unit, Royal Hallamshire Hospital, University of Sheffield, Sheffield, UK

ARTICLE INFO

Keywords:
Transient loss of consciousness
Epilepsy
Functional seizures
Classification
Natural language processing
Speech analysis

ABSTRACT

Objective: A clinical decision tool for Transient Loss of Consciousness (TLOC) could reduce currently high misdiagnosis rates and waiting times for specialist assessments. Most clinical decision tools based on patient-reported symptom inventories only distinguish between two of the three most common causes of TLOC (epilepsy, functional /dissociative seizures, and syncope) or struggle with the particularly challenging differentiation between epilepsy and FDS. Based on previous research describing differences in spoken accounts of epileptic seizures and FDS seizures, this study explored the feasibility of predicting the cause of TLOC by combining the automated analysis of patient-reported symptoms and spoken TLOC descriptions.

Method: Participants completed an online web application that consisted of a 34-item medical history and symptom questionnaire (iPEP) and spoken interaction with a virtual agent (VA) that asked eight questions about the most recent experience of TLOC. Support Vector Machines (SVM) were trained using different combinations of features and nested leave-one-out cross validation. The iPEP provided a baseline performance. Inspired by previous qualitative research three spoken language based feature sets were designed to assess: (1) formulation effort, (2) the proportion of words from different semantic categories, and (3) verb, adverb, and adjective usage.

Results: 76 participants completed the application (Epilepsy = 24, FDS = 36, syncope = 16). Only 61 participants also completed the VA interaction (Epilepsy = 20, FDS = 29, syncope = 12). The iPEP model accurately predicted 65.8 % of all diagnoses, but the inclusion of the language features increased the accuracy to 85.5 % by improving the differential diagnosis between epilepsy and FDS.

Conclusion: These findings suggest that an automated analysis of TLOC descriptions collected using an online web application and VA could improve the accuracy of current clinical decisions tools for TLOC and facilitate clinical stratification processes (such as ensuring appropriate referral to cardiological versus neurological investigation and management pathways).

1. Introduction

Transient Loss of Consciousness (TLOC) is a time-limited loss of awareness characterised by abnormal motor control, loss of responsiveness, amnesia, and a complete recovery. Over 90 % of TLOC presentations are explained by epilepsy, functional/dissociative seizures (FDS) or syncope [5]. A thorough analysis of the medical history by an expert is currently the most effective differential diagnostic method [21] because patients are typically asymptomatic on presentation and investigations after the event are of limited value. However, patients with TLOC usually present in non-expert primary or emergency care settings, and at least 20 % % of patients are initially misdiagnosed [32]. This means that many undergo inappropriate investigations and receive ineffective treatment. To date there are no reliable technical aids available that would be capable of differentiating reliably between the three most frequent causes of TLOC [30].

A symptom and medical history questionnaire called the iPEP has demonstrated promise for a three-way classification of patients presenting with TLOC [31], although its clinical effectiveness remains unproven at this stage. Modelling of responses of patients with chronic TLOC disorders using a Random Forest classifier trained with data captured by this 34-item patient questionnaire suggested that an overall

[*] Corresponding author.
E-mail address: n.pevy@sheffield.ac.uk (N. Pevy).

https://doi.org/10.1016/j.seizure.2023.11.022
Received 3 July 2023; Received in revised form 29 October 2023; Accepted 17 November 2023
Available online 7 December 2023

accuracy of 78.3 % might be achieved with this tool (83.8 % syncope, 81.5 % epilepsy and 67.9 % FDS). Furthermore, combining patient reported symptoms with witness observations increased the accuracy in the modelling to 86 % % (100 % syncope, 85.7 % epilepsy and 75 % FDS). While the model identified all instances of syncope, it performed less effectively at differentiating between epilepsy and dissociative seizures, suggesting the need for further research to improve this particular differential diagnosis.

One method which could improve the diagnostic potential of symptom inventories involves the automated analysis of language. Previous qualitative research using Conversation Analysis (CA) has identified a range of interactional differences between how people with epilepsy or FDS describe their seizure experience. People with epilepsy typically focus on their subjective seizure experience, include extensive details about their seizure symptoms, and try to reconstruct periods of loss of awareness [27,28]. In doing so, they exhibit more formulation effort in the description of their seizure symptoms (characterised by hesitations, repetitions, restarts, and reformulations) [27,28]. In contrast, people with FDS are more likely to focus on what they do not know about their seizure manifestations by making more complete negations (e.g. "I can't remember anything") [27,28] while focusing on the situations in which their seizures occur or the consequences of their seizures. Additional differences between the communication styles of these two patient groups have been described in their use of metaphoric conceptualizations [22], labels for their chief complaint [23], and references to third parties [26]. These findings - originally made in German and English speakers - have since been replicated in patients speaking a range of different languages and qualitative analysis of such observations by experts in CA allowed raters to predict the medical diagnosis of epilepsy or FDS with accuracies ranging between 80 and 90 % [3,4,6,16,24,33]. These findings demonstrate clear and clinically relevant differences in the speech and language used by people with epilepsy or FDS when talking about their seizure experiences. However, the reliance on experts in CA limits the scalability of this diagnostic approach.

A fully automated pipeline capable of capturing and analyzing spoken descriptions of TLOC could address this limitation and would allow combination with questionnaire-based clinical decision tools. We have previously demonstrated the feasibility of differentiating between epilepsy and FDS by automatically applying two linguistic feature sets to transcripts of doctor-patient interactions: Firstly we showed that a feature set designed to capture differences in formulation effort (e.g. hesitations, repetitions, and how patients pause) were capable of predicting the diagnosis with an accuracy of 71 % [19]. Secondly, we found that semantic differences in patients' language measured using 21 categories from the Linguistic Inquiry and Word Count (LIWC) [18] were capable of distinguishing between the two diagnoses with an overall accuracy of approximately 78–81 % [20]. While our studies with these feature sets demonstrated the feasibility of an automatic diagnostic approach based on the analysis of spoken language, they are not an exhaustive exploration of all features of potential differential diagnostic value which would be suitable for automatic detection. For example, the previously used feature sets were not specifically optimised to examine differences in the descriptions of subjective seizure symptoms. It is also unclear from previous work whether the diagnostic performance would be maintained if patients' spoken descriptions of their seizures were not sampled from interactions with clinicians but from the interaction with a computer-presented virtual agent (VA). Lastly is uncertain whether linguistic features could improve the diagnostic accuracy of questionnaire-based methods.

1.1. Aim

The aim of this research paper is to explore whether the automated analysis of spoken seizure descriptions can improve the differential diagnostic performance of the iPEP questionnaire in the assessment of patients with TLOC. Firstly, the performance of the formulation effort and LIWC features from previous research were evaluated on seizure descriptions collected using a VA. Secondly, the performance of an additional feature set based on the usage of verbs, adjectives, and adverbs was evaluated to detect differences in subjective symptom descriptions. Thirdly, we evaluated whether the baseline performance of the iPEP could be improved by the inclusion of these language features.

2. Method

2.1. Recruitment

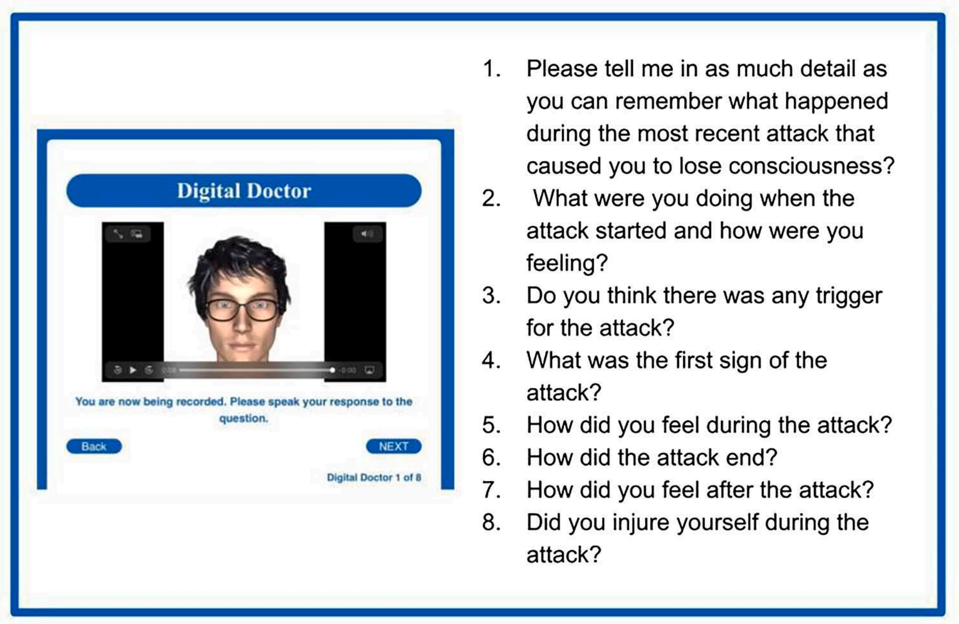
Patients received information about the study alongside appointment letters for the seizure and syncope clinics at the Royal Hallamshire Hospital. Recruitment targeted individuals over the age of 16 with a diagnosis or epilepsy, FDS, or syncope. Information about the study was also posted through various communication channels by the following charities supporting individuals with TLOC: STARS, Epilepsy Action, FNDHope, FNDAction, Epilepsy Sparks, and the Shape network supported by Epilepsy Research UK. Potential participants completed a "consent-to-contact" form and were approached by a member of the research team. The study therefore used a convenience sample. Patients who had agreed to complete the study procedure were encouraged to recruit witnesses of their seizures to provide additional information via the same online platform used for patient recruits. However, as responses from witnesses were only available for 34 % only patient responses were used for the present analysis because the sample size for witness responses was insufficient to train a machine learning model. The Leicester South Ethics Committee reviewed and granted ethical permission for this research (REC reference: 20/EM/0106).

Participants were either diagnosed using video-EEG, clinical assessment by a trained epileptologist, or both. The diagnosis was confirmed using the medical record for participants recruited via the Royal Hallamshire Hospital, whereas self-disclosure was used for the additional participants who were recruited externally.

Participants received a link to an online web application that consisted of a demographic and seizure history questionnaire, the iPEP, and an interaction with an unresponsive virtual agent (VA). VAs have been used for the remote collection of spoken descriptions of health for research studies using similar applications, for example an application designed to detect signs of dementia [15]. The iPEP asked patients 42 questions about their medical history and symptoms [31]. Witnesses (if available) were asked 10 questions about what they had observed during the attack [31]. The questions asked by the VA were designed to mirror the questions typically asked during routine epilepsy clinic consultations. In order to interact with the VA, participants were instructed to play short videos showing the VA posing the question and then responding to the question by speaking to the VA as if they were speaking to a clinician (Fig. 1). Spoken responses were recorded using the computer's inbuilt microphone. There were eight questions for patients and four questions for witnesses (if available). The first question asked participants to provide as much detail about the most recent attack as possible, and the follow-up question probed for additional details, for example what was happening before, during and after they lost consciousness (Fig. 1).

2.2. Dataset

A total of 78 patients were recruited to the study. All participants completed the questionnaire, but only 61 completed the interaction with the VA due to technological and time constraints. A breakdown of the demographic and seizure history information is given in Table 1. Patients with FDS have a high seizure frequency and number of hospitalisations, reflecting findings in other research samples collected in a similar context [25], although the severity of FDS captured in our study may be related to the fact that some of the patients recruited had been referred to a specialist seizure clinic.



Digital Doctor

You are now being recorded. Please speak your response to the question.

1. Please tell me in as much detail as you can remember what happened during the most recent attack that caused you to lose consciousness?
2. What were you doing when the attack started and how were you feeling?
3. Do you think there was any trigger for the attack?
4. What was the first sign of the attack?
5. How did you feel during the attack?
6. How did the attack end?
7. How did you feel after the attack?
8. Did you injure yourself during the attack?

Fig. 1. A still shot of the virtual agent from the web application and the eight questions that were asked. The study uses a prototype virtual agent to allow the feasibility of this approach to be tested. If the application were to be taken forward, the design of the application and virtual agent would be greatly improved.

Table 1
A breakdown of the demographic and seizure history information for each diagnostic group of patients who completed the application. The questions were presented in different formats: {a} free text box, {b} choice between four options (None, up to 5, 5:50, more than 50), {c} choice between four options (Never, Once, Up to 5, and more than 5), and {d} indicate "yes" or "no" (the percentage of "yes" responses is reported per group).

	Epilepsy	FDS	Syncope
How old were you when you had your first seizure? {a}	43 (15.5)	36 (27.1)	55 (23.2)
How many years have you been having seizures for? {a}	32.4 (20.7)	31 (15.6)	49.8 (25)
How many seizures have you had in the last year? {b}	Up to 5	More than 50	Up to 5
How many times have you been to hospital due to a seizure? {c}	Up to 5	Up to 5	Once
Have you been to intensive care due to a seizure? {d}	4.2 %	11.1 %	0 %
Do you have a family history of seizures? {d}	16.7 %	13.9 %	37.5 %
Gender (Male)	33.3 %	13.9 %	37.5 %

Most participants were White/British (89.7 %). Most patients (52.7 %) had at least one degree (equivalent of at least 16 years in education), 13.5 % had A-Level or equivalent (equivalent of 13 years in education), 13.5 % had GCSE (equivalent of 11 years in education), 2.7 % had no educational qualifications, and 4 participants did not specify.

2.3. Feature extraction

The audio recordings were manually transcribed and pre-processed to remove punctuation, convert all text to lowercase, expand contractions, and extract patient only talk. Three independent feature sets were extracted and evaluated for the language analysis.

The first set of features was designed to capture formulation effort and extracted using a manually defined script. Seven features measured the number of hesitations, repetitions, proportion of key words associated with uncertainty (uncertainty was assessed using the "nonfluency" category from the LIWC application [18]), and the frequency, average, and total duration of pauses. The group differences for these features have been detailed in a previous research paper [19]. Pauses were detected using the Web RTC Voice Activity Detector by Google.

The second set of features was designed to capture differences in the semantic content using the LIWC application (2022) [18]. Nine semantic categories were selected ("Focus present tense", "Emotional tone", "Tentativeness", "Quantifiers", "Reward", "Social", "Affect", "We" and "He/She") because these features had emerged as having the largest impact on classification accuracy between people with epilepsy or FDS when evaluated on a different dataset [20].

The third feature set measured the frequency of specific adjectives, adverbs, and verbs to detect differences in the description of subjective symptoms and the action that surrounded the period of unconsciousness [27,28]. The text was lemmatised and all verbs, adjectives, and adverbs were identified using Spacy [10]. The Term Frequency Inverse Document Frequency (TFIDF) vectoriser from Scikit Learn library in Python [17] was used to convert the verbs, adjectives, and adverbs into vector representations. TFIDF is a simple and efficient method of representing a document as a set of terms that can be easily interpreted [1].

This analysis only focused on single words that were included in a minimum of three documents and no more than seven. The algorithm was applied to the training data for each fold of the leave-one-out cross validation procedure, and the maximum number of features (N) was determined by evaluating the predictive performance of different values (10, 20, 50, 100) using a "nested" fivefold cross validation that was restricted to the training data [29].

Missing values within the iPEP questionnaire were imputed using K-Nearest Neighbour (KNN). A KNN model was trained for each model and used to predict the missing responses.

2.4. Data augmentation

Data augmentation was performed on the training data for each cross-validation fold using a method called Adaptive Synthetic sampling (ADASYN) [9] to balance the samples by up-sampling the number of samples in the epilepsy and syncope groups.

2.5. Classification

The classification performance of the iPEP questionnaire and each language feature set was evaluated separately using a Support Vector Machine with an RBF kernel (chosen because it exhibited the greatest and least unstable performance in a previous study) [20]. All models were trained to classify cases to one of three diagnoses (epilepsy / FDS / syncope) using nested leave-one-out cross validation [29]. Given that the aim was to explore whether the language features can improve the classification performance of the iPEP, two methods of integrating the iPEP and language features were evaluated. The first method (henceforth named ”all features'') involved training a single model using all features (iPEP and language features) and all diagnostic groups. Given that the iPEP exhibited particularly high sensitivity for syncope [31], the second method (henceforth called ”stacking”), used a two tier approach (Fig. 2). Firstly, a model trained using the iPEP responses was used to make predictions. This iPEP model was used to make the final decision if the prediction was syncope or if the participant did not go on to interact with the VA having completed the iPEP. If this iPEP model made a prediction of epilepsy or FDS, the data was passed into a second stage analysis whereby three separate language analysis models were used to make predictions using each of the language analysis feature sets (formulation effort, LIWC, and TF-IDF). This means that, when the iPEP prediction was included, there were a total of four predictions for each participant who had interacted with the VA. Finally, a higher-order model was trained using a SVM, LOOCV and all four predictions. These final predictions were combined with the predictions for participants that were not included in the second stage analysis and used for evaluation.

3. Result

3.1. Binary classification between

All language feature sets were good at the binary classification between epilepsy and FDS, but the formulation effort and LIWC features exhibited the best performance with accuracies of 85.7 % (Table 2).

3.2. Three-way classification between

The model trained using the patient symptoms from the iPEP questionnaire demonstrated an accuracy of 66 % for the three-way classification between epilepsy, FDS and syncope (Table 3). The performance of the three language feature sets was dramatically reduced by the inclusion of people with syncope across all feature sets (Table 3). The formulation effort features still exhibited the best performance out of the three language feature sets with an accuracy of 59 %, although this is likely due to the high sensitivity for FDS and over-representation of FDS in the sample. Compared to the baseline performance exhibited by the iPEP features alone, the ”stacking” approach resulted in an accuracy increase of 20 % and improved the detection of people with epilepsy or FDS by 29 % and 22 %, respectively (Table 3). In contrast, training a model using the ”all features'' method reduced the performance of the iPEP.

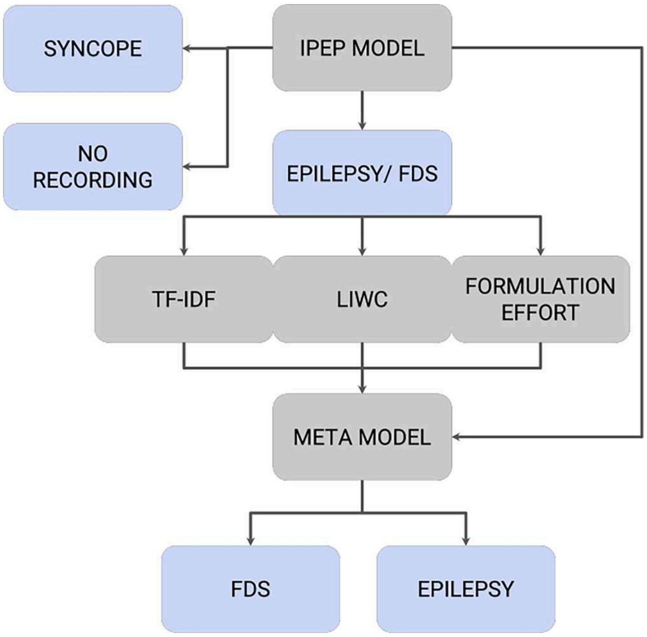


Fig. 2. A representation of the model stacking algorithm. The grey boxes represent the different machine learning models. The blue boxes represent diagnostic predictions. All predictions were used to evaluate.

Table 2
Model performance for the binary classification using each feature subset from the language analysis.

Features	Accuracy	Epilepsy	FDS
Formulation Effort	86	65	100
LIWC	86	70	97
TF-IDF	76	70	79

4. Discussion

This project investigated the feasibility of predicting the cause of TLOC using an online web application. The results demonstrate that it is possible to differentiate between epilepsy and FDS using an automated analysis of seizure descriptions produced during an interaction with a VA. Furthermore, our findings demonstrate that automated analysis of seizure descriptions can improve the performance of symptom checklist based clinical decision tools designed for the three-way classification between epilepsy, FDS, and syncope by improving the challenging differentiation between patients with epilepsy and FDS through the analysis of spoken descriptions of seizures [31]. These descriptions can be elicited by an unresponsive virtual agent. The combination of data collection using a VA and automated analyses produces a system that is more readily scalable and could be made available to aid differential diagnosis in primary and emergency departments at all hours of the day. In such non-expert settings, an automated patient stratification system could ensure that patients are referred to the correct medical specialities (e.g. neurology versus cardiology) and receive the most appropriate investigations (e.g. brain scans and EEG versus prolonged ECG and blood pressure or tilt table tests).

In contrast to the most popular machine learning approach of training a single model using all features and all classes, this research used an analytical pipeline that restricted the automated analysis of language to a downstream comparison between people with epilepsy and FDS. Our findings demonstrate that it is feasible to use a comparatively simple model to conduct the differentiation between syncope and seizures (based on a symptom checklist alone) and a more complicated model to differentiate epilepsy and FDS.

Our findings demonstrate that it is not always necessary to identify

Table 3
Model performance for the three-way classification using each feature subset. iPEP: the condensed Paroxysmal Event Profile developed by Wardrope et al. [31]. LIWC: Linguistic Inquiry and Word Count. TF-IDF: Term Frequency Inverse Document Frequency. FDS: Functional (Dissociative) Seizures.

Features	Accuracy	Epilepsy	FDS	Syncope
IPEP	66	54	72	69
Formulation Effort	59	20	100	25
LIWC	32.8	100	0	0
TF-IDF	52.5	70	45	42
All Features	59	20	100	25
Stacking	86	83	94	69

features that are highly discriminant in multi-class classification problems. Multi-class classification problems can be segmented into a multistage analysis whereby the more challenging differential diagnoses are conducted using a more complex analysis at a later stage of the analytic pipeline.

Machine learning applications in the field of medicine are often limited by the size of available data sets [12]. Small data sets hinder the identification of reliably discriminating features because it is not possible iteratively to conduct feature engineering and selection without producing bias estimates of performance [29]. This analysis has demonstrated our formulation effort and LIWC features, which have been tested in previous studies [19,20] can display high classification performance on a novel data set, which provides further evidence that these features have diagnostic utility in the differential diagnosis of epilepsy and FDS.

In contrast, the classification performance of the iPEP features was much lower than suggested by previous modelling. The accuracy reported in this study was 12.3 % less than predicted in the modelling of this questionnaire, and whereas the iPEP previously identified patients with syncope better than the other classification approaches used here, among the participants of this study, it was most accurate at predicting diagnoses of FDS. This reduction of performance may be caused by the modest sample size in the present study (especially of the syncope subgroup). However, it may also be that the binary response format of the iPEP questionnaire used in patients with a less chronic TLOC disorder performs less well in the diagnostic classification task than the 5-point likert scales which were used in the original version of this questionnaire [25,31]. Therefore, further validation work should be conducted to explore whether the performance metrics observed in the modelling of the iPEP can be replicated when the questionnaire is administered in a binary format because the analysis outlined in this paper is dependent on the iPEP reliably identifying most individuals with a diagnosis of syncope.

This study was the first paper to explore the feasibility of predicting the cause of TLOC based on the usage of verbs, adjectives, and adverbs. These features exhibited a reasonably predictive performance that supports the use of these features in future machine learning models. Overall, these findings provide further support for the feasibility of the automated differentiation between epilepsy and FDS using an automated analysis of language.

The predictive performance of each feature set was independently reduced by including individuals with syncope into the model. The reduction in accuracy may be because individuals with syncope produce spoken descriptions that are similar to the descriptions from individuals with epilepsy or FDS, making it difficult for the model to identify patterns in the features that are a reliable indicator of a single diagnosis. It is unsurprising that the performance of these two models was reduced given that these features were originally selected to discriminate between epilepsy and FDS. Furthermore, the patients with syncope who were referred to specialist clinics and recruited to this study may have had attacks bearing a closer resemblance to epilepsy than patients only assessed in primary care and emergency care settings with more typical syncope presentations. It may be that, in a study capturing patients with syncope who were never referred to specialist assessment, the recognition of syncope and overall diagnostic performance of the combination of iPEP and virtual assistant would have been better.

4.1. Limitations

The analysis used a small sample size, especially for participants with syncope. Furthermore, not all participants received a gold-standard diagnosis, for example using video-EEG. As the majority of patients in the UK receive diagnoses of epilepsy or FDS which have not been proven by video-EEG [14], our inclusion of patients whose diagnoses were not based on this diagnostic gold-standard will mean that our findings are based on a less highly selected patient population and more relevant to routine practice. However, possible inaccuracies of clinical diagnoses may have influenced the performance metrics of the models.

There has not been an exhaustive exploration of features because a separate training and testing data set is required for feature engineering and selection [29]. Future research should explore alternative features to fully optimise this classification problem, especially for individuals with syncope for whom there has been no previous research exploring language features that can aid the identification of this diagnosis.

This paper provides no indication for how well this approach would work after the inclusion of automatic speech recognition (ASR). A sufficiently large data set is required to adequately evaluate the performance of ASR because ASR models require fine-tuning using data from the target domain. Given the limited size of the sample available for this study, an ASR model would likely provide inaccurate estimates of real-world performance. Therefore, future research should aim to develop a tailored ASR system for interactions about TLOC using a larger and cleaned dataset.

The sample used in this study was not ethnically diverse because most participants were white and British. The data used to train an ASR system often uses speech from individuals who are native speakers of the target language, but these models can perform less effectively for individuals who are non-native speakers of the target language [7]. Therefore, ethnicity can have an impact on the performance of an automated analysis of language [12], and these confound variables, alongside additional confounds, should be explored more extensively in future research.

We were unable to evaluate the performance of the iPEP when witness responses were incorporated due to a limited number of participants. The performance of the iPEP has been shown to improve when witness responses are incorporated to the extent that all individuals with syncope are correctly identified [31]. The fact that witness accounts were lacking from a proportion of the participants of our study reflects the clinical reality that many TLOC patients present to medical services when no witnesses are available.Therefore, future research should explore the impact of witness contributions, including whether an automated analysis of witness descriptions of the most recent attack can aid the differential diagnosis given that witness descriptions are vital in the differential diagnostic process [13].

Last but not least – while information about user satisfaction with the iPEP and VA system was collected in the form of patient and observer interviews (which will be analyzed and reported separately), we did not seek any views from clinicians about their faith in classifications generated by an AI process or about how the VA may be deployed clinicall in the future.

5. Future research

Future research should explore methods of improving the machine learning algorithm. Increasing the sample size would allow further development and improvement of the model. Alternative or additional features that are dependent on sample size could be developed, for example large language models could be used to develop features that are potentially effective predictors [8]. Other sources of data could be integrated into the model if it is possible to automate the analysis of the data, for example witness descriptions of the attack, body language [2], and home recordings of the seizure [11]. The design of the application can be further improved based on feedback from patients and clinicians, which could potentially include newer technology advanced in interactive speech technology to improve the interaction between the patient and VA.

6. Conclusion

This paper has explored the feasibility of predicting the cause of TLOC using an online patient symptoms and witness observation questionnaire (iPEP) and the automated analysis of spoken descriptions of

TLOC. We demonstrated that it is possible to improve the challenging differentiation between people with epilepsy or FDS using the automated analysis of seizure descriptions. However, increases in performance were only achieved when the automated analysis of language was restricted to people with epilepsy and FDS. These findings demonstrate the feasibility of using this method to improve the differential diagnosis, but future research can improve upon this research by exploring whether the predictive performance of the version of the iPEP as administered through the online web application can be improved by training a machine learning model using a larger sample size, identifying linguistic features that are useful for identifying individuals with syncope, creating an ASR system that is tailored towards descriptions of TLOC, and identifying and mitigating confounding variables. Finally, it is important to evaluate the acceptability of the approach from the perspective of users to ensure this is a clinical decision tool that patients and witnesses would be prepared to use.

Ethics in publishing

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Declaration of Competing Interest

None of the authors have any conflict of interest to disclose

Acknowledgments

This research project was funded by Epilepsy Research UK, grant number 160296-1.

References

[1] Alsmadi I, Gan KH. Review of short-text classification. Int J Web Inf Syst 2019;15 (2):155–82.

[2] Alzahrani F, Mirheidari B, Blackburn D, Maddock S, Christensen H. Eye blink rate based detection of cognitive impairment using in-the-wild data. In: Proceedings of the 9th international conference on affective computing and intelligent interaction (ACII). IEEE; 2021. p. 1–8.

[3] Beghi M, Cornaggia I, Diotti S, Erba G, Harder G, Magaudda A, Laganà A, Vitale C, Cornaggia CM. The semantics of epileptic and psychogenic nonepileptic seizures and their differential diagnosis. Epilepsy Behav 2020;111:107250.

[4] Biberon J, de Liège A, de Toffol B, Limousin N, El-Hage W, Florence AM, Duwicquet C. Differentiating PNES from epileptic seizures using conversational analysis on French patients: a prospective blinded study. Epilepsy Behav 2020;111: 107239.

[5] Brignole M, Moya A, De Lange FJ, Deharo JC, Elliott PM, Fanciulli A, Fedorowski A, Furlan R, Kenny RA, Martín A, Probst V. Practical Instructions for the 2018 ESC Guidelines for the diagnosis and management of syncope. Eur Heart J 2018;39(21):e43–80.

[6] Cornaggia CM, Gugliotta SC, Magaudda A, Alfa R, Beghi M, Polita M. Conversation analysis in the differential diagnosis of Italian patients with epileptic or psychogenic non-epileptic seizures: a blind prospective study. Epilepsy Behav 2012;25(4):598–604.

[7] Cumbal R, Moell B, Águas Lopes JD, Engwall O. You don't understand me!": comparing ASR results for L1 and L2 speakers of Swedish. In: Proceedings of the Interspeech. 2021; 2021.

[8] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

[9] He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the IEEE international joint conference on neural networks (IEEE world congress on computational intelligence. Ieee; 2008. p. 1322–8.

[10] Honnibal, M., & Montani, I. (2017). spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[11] Karakas C, Modiano Y, Van Ness PC, Gavvala JR, Pacheco V, Fadipe M, Thanaviratananich S, Alobaidy AM, Purohit A, Fussner S, Chen DK. Home video prediction of epileptic vs. nonepileptic seizures in US veterans. Epilepsy Behav 2021;117:107811.

[12] Latif S, Qadir J, Qayyum A, Usama M, Younis S. Speech technology for healthcare: opportunities, challenges, and state of the art. IEEE Rev Biomed Eng 2020;14: 342–56.

[13] Malmgren, K., Reuber, M. and Appleton, R., 2012. Differential diagnosis of epilepsy. Oxford textbook of epilepsy and epileptic seizures, pp.81–94.

[14] Mayor R, Smith PE, Reuber M. Management of patients with nonepileptic attack disorder in the United Kingdom: a survey of health care professionals. Epilepsy Behav 2011;21(4):402–6.

[15] Mirheidari B, Blackburn DJ, Harkness K, Walker T, Venneri A, Reuber M, Christensen H. An avatar-based system for identifying individuals likely to develop dementia. In: Proceedings of the Interspeech; 2017. p. 3147–51. 2017ISCA.

[16] Papagno C, Montali L, Turner K, Frigerio A, Sirtori M, Zambrelli E, Chiesa V, Canevini MP. Differentiating PNES from epileptic seizures using conversational analysis. Epilepsy Behav 2017;76:46–50.

[17] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[18] Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001, 71. Mahway: Lawrence Erlbaum Associates; 2001. p. 2001.

[19] 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;91: 141–5.

[20] 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 Behav 2023;143:109217.

[21] Plug L, Reuber M. Making the diagnosis in patients with blackouts: it's all in the history. Pract Neurol 2009;9(1):4–15.

[22] Plug L, Sharrack B, Reuber M. Seizure metaphors differ in patients' accounts of epileptic and psychogenic nonepileptic seizures. Epilepsia 2009;50(5):994–1000.

[23] Plug L, Sharrack B, Reuber M. Seizure, fit or attack? The use of diagnostic labels by patients with epileptic or non-epileptic seizures. Appl Linguist 2010;31(1):94–114.

[24] Reuber M, Monzoni C, Sharrack B, Plug L. Using interactional and linguistic analysis to distinguish between epileptic and psychogenic nonepileptic seizures: a prospective, blinded multirater study. Epilepsy Behav 2009;16(1):139–44.

[25] Reuber M, Chen M, Jamnadas-Khoda J, Broadhurst M, Wall M, Grünewald RA, Howell SJ, Koepp M, Parry S, Sisodiya S, Walker M. Value of patient-reported symptoms in the diagnosis of transient loss of consciousness. Neurology 2016;87 (6):625–33.

[26] Robson C, Drew P, Walker T, Reuber M. Catastrophising and normalising in patient's accounts of their seizure experiences. Seizure 2012;21(10):795–801.

[27] Schwabe M, Howell SJ, Reuber M. Differential diagnosis of seizure disorders: a conversation analytic approach. Soc Sci Med 2007;65(4):712–24.

[28] Schwabe M, Reuber M, Schondienst M, Gulich E. Listening to people with seizures: how can linguistic analysis help in the differential diagnosis of seizure disorders? Commun Med 2008;5(1):59.

[29] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14(11):e0224365.

[30] Wardrope A, Newberry E, Reuber M. Diagnostic criteria to aid the differential diagnosis of patients presenting with transient loss of consciousness: a systematic review. Seizure, 2018;61:139–48.

[31] Wardrope A, Jamnadas-Khoda J, Broadhurst M, Grünewald RA, Heaton TJ, Howell SJ, Koepp M, Parry SW, Sisodiya S, Walker MC, Reuber M. Machine learning as a diagnostic decision aid for patients with transient loss of consciousness. Neurol Clin Pract 2020;10(2):96–105.

[32] Xu Y, Nguyen D, Mohamed A, Carcel C, Li Q, Kutlubaev MA, Anderson CS, Hackett ML. Frequency of a false positive diagnosis of epilepsy: a systematic review of observational studies. Seizure, 2016;41:167–74.

[33] Yao Y, MA W, Markus R, LU Q, Huang Y, Zhou X, Dou W, WU L, Yao X, Liu L, Yuan Y. Conversation analysis in differential diagnosis between epileptic seizure and psychogenic nonepileptic seizure. Chin J Neurol 2017:266–70.