University of Sheffield

Improving Automatic Speech Recognition to help the prediction of Transient Loss of Consciousness Consultations

Baiyang Qu

Supervisor: Heidi Christensen

A report submitted in fulfilment of the requirements
for the degree of MSc in Advanced Computer Science

in the

Department of Computer Science

September 13, 2023

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Baiyang Qu

Signature: Baiyang Qu

Date: 26/07/2023

Abstract

Transient loss of consciousness (TLOC) refers to a brief and temporary episode during which a person loses consciousness or awareness. TLOC has a significant impact on the daily lives of many people. The majority of TLOC cases are caused by one of three medical conditions: epilepsy, functional (dissociative) seizures (FDS), or syncope (Pevy, 2023). There are many diagnostic and discriminatory methods that have emerged over the decades, but none of them are completely reliable. Up to 20% of people receive an incorrect diagnosis at the outset. With the development of speech technology, the exploration of using conversation analysis to aid TLOC cause detection has emerged as a fascinating area of interest in recent times. This project aims to design and implement a complete machine learning pipeline for automating the diagnosis of the cause of TLOC using patient audio input. The two feature sets were designed to assess formulation effort or identify semantic distinctions among patients with one of three health conditions. As the baseline model of this project, the two feature sets were integrated with machine learning models, resulting in diagnostic accuracy rates of 54.5% and 59.2%, respectively. When this method is applied to binary classification, epilepsy and FDS, the accuracy rate will increase to 63.9% and 74.1%. Furthermore, Bert embedding with a Long Short-Term Memory (LSTM) structure is also applied to this task.

To automate the TLOC consultation process, various automatic speech recognition (ASR) models are used to generate transcripts of patient's speech. Then, the performance of two large-scale pre-trained ASR models, ESPnet and Whisper, is compared on the real-life dataset. The Whisper-medium model, with a 32.84% WER performance, outperformed ESPnet's 53.85% WER performance as the more appropriate ASR model for the task.

The designed differential diagnosis pipeline achieved an accuracy of 77.6% in distinguishing between two types of seizure: epilepsy and FDS, surpassing the accuracy of 62.4% in the three-way classification (epilepsy, FDS, and syncope) on real-world data. This also indicates that the pipeline implemented in this project can offer effective guidance for diagnosing the causes of TLOC.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Professor Heidi Christensen and Dr. Nathan Pevy. Thank you for leading me into this enjoyable topic that can benefit our society. I would not have been able to complete my dissertation project without your generous support and guidance. Besides, I really appreciate the important advice you provided me at crucial stages.

Also, I would like to thank my parents and April. You always give me love and support not only in my master's study but also in my life over the years.

I will never forget the moments staying at the University of Sheffield. This year has been an extremely valuable and colourful time for me. The mixing of different cultures has made me more receptive to the diverse world. Yes, this dissertation marks the end of my master's program, but also the beginning of a new chapter in my life.

Contents

List of Figures

List of Tables

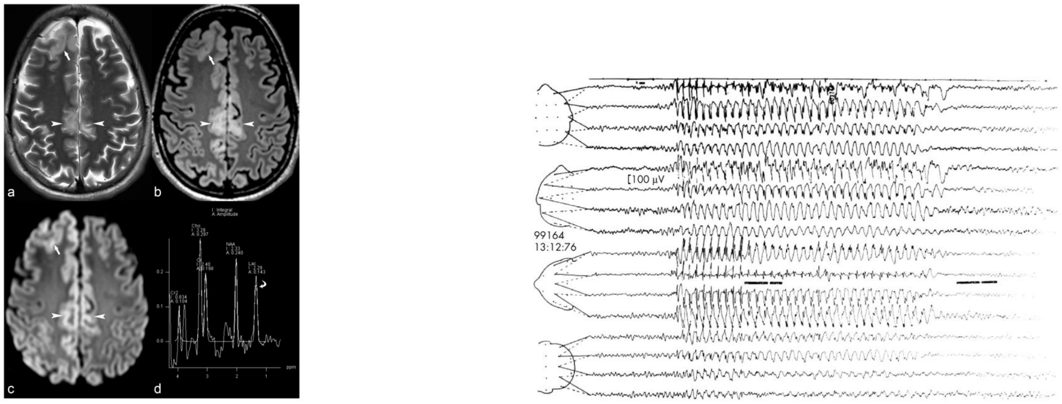
Chapter 1

Introduction

According to NICE, transient loss of consciousness (TLOC) impacts as many as half of the UK's population (O'Flynn, 2011). TLOC is common in primary care. TLOC refers to a momentary state of unawareness marked by amnesia, unusual motor control, unresponsiveness, and a brief duration without apparent causes, followed by a complete recovery (Dijk et al., 2009).

In the majority of cases (over 90%), TLOC can be attributed to three primary medical conditions: epilepsy, functional (dissociative) seizures (FDS), and syncope. Accurately differentiating between the various causes of TLOC is of utmost importance to ensure Effective management and the recognition of individuals susceptible to morbidity or mortality due to various underlying health conditions (Wardrope et al., 2018). Distinguishing between the diverse causes of blackouts is a significant and challenging clinical issue. The success of treating epilepsy relies heavily on the initial diagnosis, making the correct categorization of the issue essential for effective therapeutic outcomes (Plug L., 2009). However, there is a notable prevalence of misdiagnoses when it comes to identifying the causes of TLOC, ranging from 20% to 30% (Chadwick and Smith, 2002). Diagnostic delay is particularly prevalent in Psychogenic Non-Epileptic Seizures (PNES), with patients experiencing a considerable time gap from initial manifestation to diagnosis, sometimes spanning several years. These delays can expose patients to an increased risk of harm and, in some cases, even death.(Reuber et al., 2016). In cases where individuals receive an incorrect initial diagnosis, they may undergo irrelevant tests, leading to delays in accessing appropriate treatment.

TLOC is linked to a variety of symptoms, which may provide varying levels of indication regarding a specific cause but cannot solely establish a differential diagnosis. While there are examinations designed to identify the underlying cause of TLOC, many of these tests can only identify one of the three TLOC causes and have limited diagnostic value during periods between TLOC episodes (Pevy, 2023). Doctors may conduct a comprehensive analysis of a patient's medical history and the details surrounding their TLOC episode to determine the cause(Plug L., 2009). Magnetic Resonance Imaging (MRI)(Zentner et al., 1999) and electrocardiogram(NICE, 2023) are also used to detect causes by measuring the structural abnormalities of the brain and the electrical activity of the heart.



(a) MRI of a 33-year-old female with myoclonic epilepsy (Zentner et al., 1999)

(b) Electrocardiogram of a seizure patient during a typical absence seizure(Smith, 2005)

Figure 1.1: Examples for MRI and electrocardiogram methods

However, the use of qualitatively evaluated methods may have limitations as they can be influenced by subjective factors, which may reduce their accuracy and applicability. Clinicians in emergency settings face the difficulty of lacking established and validated diagnostic criteria for distinguishing among the prevalent causes of TLOC in patients. (Wardrope et al., 2018) . To discover an effective diagnostic criterion for TLOC, clinicians can utilize a clinical decision tool that combines various clinical variables linearly, such as medical history and patient information. This tool helps them make diagnostic decisions and facilitates the handling of extensive clinical data(Stiell and Bennett, 2007). Compared to the classical linear method, non-linear machine learning methods based on The Random Forest algorithm were proposed to improve the classification accuracy of TLOC. For example, the iPEP procedure achieved a precise classification for 78.3% of patients (83.8% for syncope, 81.5% for epilepsy, and 67.9% for functional disorders) using solely patient data (Wardrope et al., 2019).

However, it showed there was still scope for improvement, so conversation analysis (CA) was incorporated into the clinical decision tool as a reliable detecting feature. This method involves analyzing recorded conversations describing their experience of TLOC at a micro-level to gain insights into the dynamics of human communication(Jefferson, 1985). More specifically, this approach involves posing a sequence of inquiries to patients about their anticipations for the consultation and motivating them to recount details about their initial, most severe, and most recent seizure episodes. Subsequently, the transcripts of these interactions will be analyzed to identify linguistic and interactional variations. According to a recent research, useful features such as formulation effort or semantic differences can be extracted from the audio files and the transcript of recording(Pevy, 2023). These features allow for a multifaceted analysis of speech and transcripts. For instance, one of the functions of formulation effort is to assess how much information patients give to describe their onset experience. Early clinical research shows that patients with epilepsy are more likely to give more details about how they felt and tried to describe what happened before and after the period when they were unconscious. They also tried to piece together what happened while they were unconscious in order to create a clear story of the events. Conversely, people with FDS often confused the time they were unconscious with the seizure itself. They made general statements, indicating that they didn't know what happened during that period (Schwabe et al., 2007).

In the context of the development of the CA method, there has been a growing interest in employing natural language processing and related machine learning algorithms for TLOC diagnosis through the analysis of patients' speech recordings in recent time. In the process of automated analysis of patient conversations, transcripts of speech files are required. In this way, a powerful and robust automatic speech recognition (ASR) system is vital and essential. In a previous TLOC causes detection research, the open-source Kaldi application and Librispeech was used to train the ASR system which achieved a word error rate of 39.28% (Pevy, 2023). Clearly, this figure indicates that there is still room for improvement. Taking inspiration from automatic diagnosis on Alzheimers disearse (AD) (Liu et al., 2021), the utilisation of more advanced ASR approaches, exhibiting superior performance, can enhance the accuracy of medical diagnosis predictions. Two potential ways for improving ASR system performance in this particular task are advanced model architectures and enhanced generalisation. More advanced model structures can better capture deep-seated features, while enhanced generalisation allows for better adaptation to specific domain applications. Subsequent experiments in this project explore and discuss these two optimisation approaches separately.

This project explores the potential of incorporating specialised feature extraction techniques with an advanced ASR system and statistical machine learning methods to enhance the accuracy of TLOC cause detection.

1.1 Aims and Objectives

The aim of this project is to explore methods for helping TLOC consultation through the analysis of speech and text. This involves the integration of specialized feature extraction, machine learning classification models, and advanced pre-trained ASR models into complete pipelines, as we experiment with different combinations to find the optimal solution.The objectives of the project are:

1. Extracting acoustic and linguistic features based on formulation effort and semantic categories.

2. Training various traditional machine learning models with the extracted features to classify the causes of TLOC, and selecting the most suitable model for this task.

3. Exploring the feasibility of combining BERT embeddings with neural network Long Short-Term Memory (LSTM) for helpin TLOC consultation.

4. Exploring the suitability of various pre-trained large-scale ASR models (ESPnet and Whisper) for this task.

5. Developing a full machine learning pipeline to help with predictions about the cause of TLOC and gain insights into the impact of various components on performance.

1.2 Overview of the Report

The report of this project is organised as the following chapters:

Chapter 1 presented the background and purpose of this project.

Chapter 2 contains a literature review of previous work to utilize machine learning algorithms in TLOC causes detection and similar works that aimed to develop a advanced ASR system for this application. The chapter also introduces some important algorithms used in this project.

Chapter 3 outlines the project's methodology, including the source of dataset, preprocessing steps, model architecture, and the implementation details.

Chapter 4 explains the experimental design and processes carried out, along with the discussion of results obtained from various experiments.

Chapter 5 serves as the project's conclusion, summarising the key findings of this project.

Chapter 6 identifies the potential improvements and opportunities for future work.

Chapter 2

Literature Survey

In this section, clinical conditions relevant to this research will be reviewed. The distinctions and connections among various causes of TLOC will be explored. A comparison will be made between traditional approaches and emerging machine-learning methods for analyzing TLOC causes. Furthermore, a deeper analysis will be conducted on the advanced algorithms and structures involved in the latest methodologies.

Firstly, the definition of TLOC, its prevalence in society, and its societal impact are presented. Furthermore, a discussion on the conditions under which TLOC occurs and its effects on the human body will be conducted.

Next, the three most probable causes and their distinctions and connections will be discussed. Subsequently, an in-depth exploration will be conducted on how patients' language usage is influenced by different causes of TLOC, and their expressions during conversations about symptoms or the onset of the condition, such as formulation effort. Based on these connections and distinctions, the approaches for determining the causes of TLOC in clinical settings will be further reviewed, enabling a comprehensive understanding of traditional diagnostic methods.

Following that, the methods of extracting language features that correspond to traditional machine learning models used for training will be reviewed. Additionally, the machine learning algorithms that have shown effectiveness in this healthcare domain will be explored. With the advancement of computing power, an increasing number of advanced models and architectures have become applicable in the healthcare domain. In light of this, advanced deep learning and pre-training models that have demonstrated promising performance in the TLOC field will be further reviewed. Subsequently, a detailed analysis of the structures and components involved in these models was conducted to gain deeper insights.

Finally, in order to achieve automated diagnosis in this system, ASR technology is essential for converting speech to text. Therefore, I conducted a review of the ASR technologies used in the past, as well as the current state-of-the-art ASR techniques. The breakthroughs in the ASR field enabled by the attention mechanism, significantly improving the accuracy of predictions, have been explored.

2.1 Clinical Condition

2.1.1 Introduction

A transient loss of consciousness event, commonly referred to as a ”blackout,” can be characterized as a sudden, temporary loss of consciousness followed by full recovery (Cooper and Westby, 2011). TLoC is a common occurrence in the general population. Recurrent TLoC episodes are linked to substantial morbidity, requiring heightened medical attention and resulting in significant impacts on the individual’s quality of life (O’Flynn, 2011).

TLOC is a symptom, not a disease, and its causes can be diverse (O’Flynn, 2011). However, diagnosing the root cause is frequently fraught with inaccuracies, inefficiencies, and delays, often resulting in prevalent misdiagnoses. (NICE, 2023). In a study involving patients with ”refractory seizures,” cardiological tests and tilt-table examinations were administered, revealing a diagnosis of syncope in over 30% of patients who were previously thought to have refractory epilepsy (Zaidi et al., 2000). Another finding is that it took an average of seven years after receiving an incorrect diagnosis for an accurate diagnosis to be made (Reuber et al., 2002). During this extended period, Nearly half of these patients were receiving inappropriate treatment with antiepileptic drugs.

These findings highlight the challenges in accurately diagnosing conditions like syncope and non-epileptic seizures, which can often be misidentified as refractory epilepsy. The delay in reaching an accurate diagnosis can lead to ineffective and potentially harmful treatments. Improving diagnostic processes is essential to ensure appropriate and timely management for patients experiencing these conditions. Therefore, distinguishing the root cause is crucial, as it ensures timely and accurate treatment while avoiding the waste of medical resources.

2.1.2 Three Causes of Transient Loss of Consciousness

The majority of TLOC cases arise from three primary health conditions: epilepsy, functional (dissociative) seizures (FDS), or syncope.

Epilepsy

According to the International League Against Epilepsy, epileptic seizures are described as occurrences of excessive and/or hypersynchronous activity in brain neurons, usually of short duration. An epileptic disorder is identified as a chronic neurological condition marked by recurrent epileptic seizures (Plug L., 2009). It is reported that there are approximately 7.6 patients will experience epileptic seizures within their lifetime in every 1000 people (Fiest et al., 2017). An epileptic seizure occurs when there is an abnormal increase in pathological electrical activity in the brain, surpassing a ”seizure threshold.” This happens when there’s an unevenness in brain activity, causing brain regions to sync up and produce coordinated oscillations due to too much excitation and too little inhibition. (Staley, 2015).

The origins and neuronal mechanisms behind the seizure dictate both the physical and psychological symptoms it presents. They also impact the particular neural circuits engaged in the seizure and the scope of brain activity affected by it.

FDS

(Functional) Non-epileptic seizures are episodes resembling epileptic seizures regarding changes in motion, sensation, or perception, but they do not exhibit the occurrence of abnormal electrical discharges in the brain during a seizure. These occurrences can be characterised as instances of losing control, brought about by distressing circumstances, emotions, thoughts, memories, or feelings. (Plug L., 2009).

FDS affects about 33 out of every 100,000 people (Benbadis and Hauser, 2000). It can show up as physical symptoms like shaking or stiffness and mental symptoms like not being aware of what's happening. People with FDS might also have different feelings and thoughts when it happens (Asadi-Pooya et al., 2016). Some think it happens because they saw someone else have a seizure, learned about seizures, or misunderstood their own body signals as a seizure starting (Brown and Reuber, 2016a). Stress, whether it's inside or outside of them, can set off these episodes.

Syncope

Syncope is characterized by a sudden and temporary loss of consciousness, often resulting in falling. It occurs due to transient global cerebral hypoperfusion (Brignole M., 2001), and its onset is typically rapid. However, recovery happens spontaneously, completely, and relatively quickly. Syncope can be categorized into three main types: vasovagal syncope, carotid sinus syndrome, and situational syncope (Brignole et al., 2018).

- Vasovagal syncope happens when people have less blood going back to their heart while standing up, which makes blood gather in their lower body. It can be set off by something that suddenly makes their heart rate go up while they're standing (Adkisson and Benditt, 2017).

- Carotid sinus syndrome is when a person faint because someone touches or presses on a certain spot in his neck called the carotid sinus. This touch messes up your body's reflex that controls blood pressure, causing low blood pressure and making you pass out (Adkisson and Benditt, 2017).

- Situational syncope happens when different situations can make people faint. Things like coughing, sneezing, laughing, swallowing, going to the bathroom, or peeing can cause it. (Adkisson and Benditt, 2017).

In summary, syncope is a complex condition with different types, each caused by distinct mechanisms and triggers. Understanding these classifications is crucial for accurate diagnosis and appropriate management of individuals experiencing syncope.

2.1.3 Differential Diagnosis Methods

The main goal of the TLOC diagnostic process is to check if the patient is facing the risk of death (Brignole et al., 2018). It starts when the patient seeks care from Primary or Emergency Services. To determine the appropriate referral route, patients initially receive an initial working diagnosis. However, this stage may lead to incorrect diagnoses for about 20% of patients (Xu et al., 2016). The neurologist then carefully reviews the patient's medical history, TLOC symptoms, and any video recordings. Electrocardiogram (ECG) help identify any heart-related issues (NICE, 2023). This thorough diagnostic process is crucial for accurately identifying the cause of TLOC and providing the right treatment.

The clinical characteristics of the attacks and the medical history

The most crucial approach for identifying the cause of TLOC lies in conducting a thorough assessment of the patient's medical history and the clinical features of their TLOC episodes.(NICE, 2023). The clinical features of TLOC include things that can make it more likely to happen, like certain triggers or situations. In the case of syncope, various triggers or circumstances may suggest this diagnosis, such as prolonged standing, extreme coughing, urination, defecation, physical exertion, or drug use (Colman et al., 2004). For people with "reflex" epileptic seizures, certain things like flashing lights, making decisions, reading, writing, getting startled, touching, feeling body movement, hearing sounds, eating, and balance can trigger their seizures (Y and Ritaccio, 2004). Additionally, many different things can trigger FDS, like breathing too fast, flashing lights, someone suggesting it, getting a certain kind of massage, fake injections, and more (Hingray et al., 2016). Most research focuses on identifying items that can aid in differentiating between epileptic seizures and other conditions. These kinds of clusters of factual items made a high accuracy(over 90%) prediction on distinguishing patients with generalized tonic-clonic seizures and syncope (Table 2) (Plug L., 2009).

In addition to factual information about the seizures themselves, there are instances where neurologists can directly observe patients' behaviour through home video or CCTV footage. However, some cases lack ictal video recordings (Plug L., 2009). In such situations, other factual information unrelated to seizures may be valuable in distinguishing between epilepsy and non-epileptic seizures. Even though the significance of these additional factual features isn't fully defined, noticeable distinctions exist between epilepsy and non-epileptic seizure patient groups (Table 4). While these features alone cannot determine the diagnosis, they can serve as "red flags" for experienced specialists to assess whether the diagnosis of epilepsy is justified.

It's worth mentioning that while various characteristics of TLOC can help with telling different causes apart, no single characteristic alone can always cause it. It's the combination of these features that experts use to figure out the most probable cause (Pevy, 2023). The problem with the implementation of this method is that it is difficult to record the full clinical characteristics and history of the patient in detail at the first onset, or worse, even to get an inaccurate record.

Feature in history	Non-epileptic seizures	Epileptic seizures
Onset under 10 years old	Unusual	Common
Change of semiology	Occasional	Rare
Aggravation by antiepileptic drugs	Occasional	Rare
Seizures in presence of doctors	Common	Unusual
Recurrent ”status” (seizures.30min)	Common	Rare
Multiple unexplained physical symptoms	Common	Rare
Multiple operations/invasive tests	Common	Rare
Psychiatric treatment	Common	Unusual
Sexual and physical abuse/emotional neglect	Common	Rare

Table 2.1: Seizure-unrelated details in the patients history which may be useful to diagnosis of epileptic seizures (Plug L., 2009)

Video-EEG

The most definitive method for diagnosing epilepsy or FDS is video-EGG(video telemetry). This process means a patient needs to be carefully watched either in the hospital or at home until they have a seizure(Pevy, 2023). At the same time, they record videos and EEG data while the seizure happens. This helps them see what happens during a seizure and if any behavioural changes are linked to brain activity. (Noachtar and Rmi, 2009).

Using Video-EEG is like taking a video of what happens during an episode of TLOC. It's useful for diagnosing FDS because it can show the behavioural patterns often seen in FDS, even if there are no typical electrical brain activity changes seen in epilepsy. This helps doctors tell the difference between the two conditions. (Brown and Reuber, 2016b).

Structural brain abnormalities

Neuroimaging techniques like MRI are crucial in epilepsy diagnosis. They help find issues in the brain related to epilepsy, such as shrinkage in the hippocampus, brain structure problems, blood vessel abnormalities, tumours, and damage from brain injuries (Fitsiori et al., 2019). Although Neuroimaging by itself might not give a definite epilepsy diagnosis, it helps a lot. It shows any brain structure problems that could cause seizures and guides treatment choices. So, it's a crucial part of assessing epilepsy patients (Malmgren et al., 2012).

Tilt-Table Testing

The Tilt-Table test is a way to diagnose vasovagal or orthostatic syncope by watching blood pressure and heart rate as a person goes from lying down to standing up (Kohno et al., 2018). It helps find out if someone is prone to vasovagal syncope. Also, it helps tell apart syncope from FDS, which looks like syncope but doesn't cause big changes in heart rate or blood pressure in FDS patients. (Tannemaat et al., 2013).

Conversation analysis on TLOC

Doctors in Sheffield have embarked on describing the interactions with patients with TLOC in terms of language and communication styles. This approach highlights the importance of how patients with epilepsy and non-epileptic seizures communicate with their doctors about their seizures, rather than solely concentrating on the symptoms they describe (Plug L., 2009). This method effectively avoids the challenges of incomplete recording of clinical characteristics and history and eliminates the potential inaccuracies associated with retrospective data collection, as it is based on real-time conversations between doctors and patients. The technique used to understand communication in healthcare interactions is conversation analysis (CA). It involves micro-analyzing recorded conversations to grasp the dynamics of human conversation (Jefferson, 1983). The research followed interview guidelines which included specific questions to ask the patients and let the patients freely share their experiences without interruption. Patients were asked to describe their seizure-related experience (Schwabe et al., 2008).

This analysis included factors such as patients' willingness to share information about their seizure experiences, their ability to provide coherent accounts of their seizure trajectory or specific episodes, and their use of hesitation markers, repetitions in constructing their medical historyreferred to as ”formulation effort” (Schwabe et al., 2007). These features differed significantly between patients with epilepsy and those with non-epileptic seizures (Table 2.2) (Plug L., 2009).

Aside from formulation effort, other linguistic features also play a significant role. For example, These characteristics have been successful in identifying emotions and detecting various conditions like stress, suicidal tendencies, distress, anxiety, depression, bipolar disorder, Parkinson's disease, ALS, and cognitive decline (Salekin et al., 2018). Linguistic features are usually applied to written transcripts of spoken words, examining the specific language patterns used by individuals with a particular health condition. This involves assessing the use of specific words related to different meanings, how fast someone talks, measures of linguistic complexity, and how often certain types of words are used. (Matton et al., 2019b). Further information about these conversation features is outlined in the next chapter.

In summary, the language used by individuals with epilepsy or non-epileptic seizures during interactions with neurologists exhibits distinct linguistic profiles. Skilled professionals can identify these differences and effectively utilize them to predict the diagnosis with great precision.

Summary

Although these tests can be used as methods to make differential diagnosis, many of them are limited by various factors. For instance, video-EEG, which involves monitoring patients during a seizure, may not always be feasible, especially for first-time patients. Additionally, the recorded information might be incomplete or inaccurate, affecting the diagnostic process.

Feature	Epilepsy	Non-epileptic seizure
Subjective seizure symptoms	Typically volunteered, discussed in detail	Avoided, discussed sparingly
Formulation work (e.g., pauses, reformulation attempts, hesitations, restarts)	Extensive, large amount of detail	Practically absent, very little detailing efforts
Seizures as a topic of discussion	Initiated by the patient	Initiated by interviewer
Focus on seizure description	Easy	Difficult or impossible
Spontaneous reference to attempted seizure suppression	Often made	Rarely made
Seizure description by negation (e.g., "I don't know", "I can't hear", "I can't remember")	Rarely, negation is usually contextualised (e.g., "I can remember this but I can't recall that")	Common and absolute (e.g., "I feel nothing", "I do not know anything has happened")
Description of periods of reduced consciousness or self-control	Common	Rare

Table 2.2: Overview of the main differences in how epilepsy and non-epileptic seizures are identified through interactions, topics, and language. (Plug L., 2009)

Similarly, EEG and neuroimaging techniques might not always provide reliable results, making it challenging to identify the exact cause of TLOC in some cases. Furthermore, certain methods, like serological biomarkers and tilt-table testing, may only be able to identify specific causes of TLOC, leaving other potential factors undetected.

2.2 Machine Learning Application in TLOC Diagnosis

Machine learning is a way to create computer systems that can learn and get better without someone telling them exactly what to do. They learn from data and use that learning to make predictions, like diagnosing illnesses in medicine. The most common type of machine learning used in medicine is called supervised learning, where the computer looks at input data and tries to predict specific outputs (Jordan and Mitchell, 2015).

2.2.1 iPEP

By analysing the patients clinical characteristics and history comprehensively, experienced specialised neurologists can make an accurate diagnosis even without any physical tests (Angus-Leppan, 2008). Acknowledging the possibility of precise diagnoses rooted in the patient's medical history, some useful clinical decision tools or rules are created to effectively make the diagnosis. One of the most important early methods is the initial Paroxysmal Event Profile (iPEP) procedure which is based on a non-linear machine learning method. Since each diagnostically relevant behavioural characteristic of the patient may carry different decision weights, the iPEP procedure takes these variations into account (Wardrope et al., 2020).

The application of the Random Forest algorithm to individual datasets involves a specific feature selection mechanism. From patient-only responses, researchers identified 34 items that accurately classified 78.3% of patients, achieving rates of 83.8% for syncope, 81.5% for epilepsy, and 67.9% for FDS (Pevy, 2023).

Although the early modelling research suggests that iPEP could improve diagnostic accuracy (Xu et al., 2016). There may be a reduction in the classification accuracy if the variables were dichotomised. And further improvements are needed, particularly in distinguishing between epilepsy and FDS. However, iPEP has opened the door to using machine learning for diagnostic prediction. Being inspired by this kind of comprehensive consideration of characteristics, further research has been conducted to explore the incorporation of a conversation analysis method of TLOC with machine learning methods.

2.2.2 Machine Learning with Speech

Machine learning methods can learn from various types of data. In healthcare, speech data is especially important. Many studies investigate whether we can predict diagnoses by analyzing speech automatically. They explore this possibility for different health conditions, such as cognitive decline (Liu et al., 2021).

The machine learning methods usually begin with the selection of features and data preprocessing. For some well-known methods, like supervised learning, we need to create features based on what we know about the problem. When analyzing conversations, the features commonly used can be divided into two groups: acoustic and linguistic features (Latif et al., 2020).

Acoustic features

There are three kinds of acoustic features: prosodic (related to pitch and rhythm), spectral and temporal (related to sound frequency and timing), and voice quality. For instance, some open source can measure features like pitch, formants, LPCC, MFCC, and GFCC (Eyben et al., 2010).

Linguistic features

Linguistic features examine the language patterns of people with a particular health condition. For instance, they look at how often certain words related to different meanings are used, how fast someone talks, how complex their language is, and how often they use certain types of words. These features have been studied in different research projects. (Matton et al., 2019c).

2.2.3 Classical Machine Learning Models

Among the traditional machine learning algorithms, those applied to TLOC diagnosis classification are mainly Random Forest, Support Vector Machines (SVM), Extreme Gradient Boosting (XGBOOST) and K-Nearest Neighbour (KNN).

Random Forest

The Random Forest algorithm generates multiple decision trees and consolidates their predictions to arrive at a final decision. Each tree consists of nodes that ask questions about features and generate child nodes with possible answers. These child nodes can further split the data or become leaf nodes, where the final predictions are based on the majority results from the training data. To improve performance, Random Forest reduces the similarity between trees by using random data samples and feature subsets during tree construction (Figure 2.2) (Breiman, 2001).

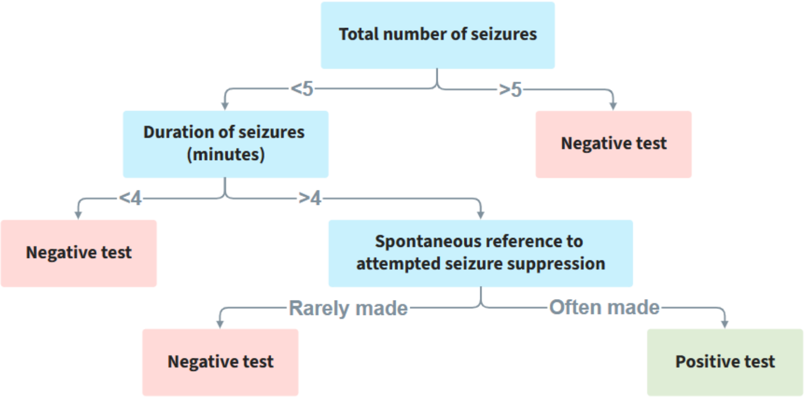


Figure 2.1: Decision tree structure

Support Vector Machine

The Support Vector Machine (SVM) is a technique that constructs a line known as a hyperplane to distinguish between different categories. It utilises this hyperplane to predict the category of new data points based on their spatial positions (Noble, 2006). The objective of SVM is to select the optimal hyperplane that maximises the distance from each category. However, in situations where data separation isn't straightforward with a straight line, particularly in intricate datasets, SVM offers solutions. It can accommodate such cases by employing a soft margin or a kernel function. The soft margin permits certain data points to reside on the incorrect side of the line without affecting the final prediction. Meanwhile, the kernel function transforms the data into a higher-dimensional space (Kim et al., 2013) to identify a more effective separation hyperplane.

K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a form of supervised machine learning, wherein it ascribes a class to fresh data based on its similarity to the training data. During the training process, each data point from the training set is plotted in a multi-dimensional space, contingent on the number of features. When we wish to evaluate the algorithm's performance, we also map the new data into the same space (Altman, 1992). Subsequently, we calculate the distances between the new data point and all other data points in the training set using the straightforward Euclidean distance method.

Following this, KNN identifies the K closest training data points to the new data point, with "K" being a predetermined number that specifies how many nearby points to take into account. KNN then determines the most prevalent class among these K nearby training data points and predicts it as the class for the new data point. One of KNN's advantages lies in its ability to handle non-linear relationships between features and classes since it doesn't rely on linear models or hyperplanes for making predictions.

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBOOST) is a powerful and popular gradient-boosting algorithm used in machine learning., it improves prediction accuracy by gathering the forecasts of many weaker learners, often in the form of decision trees, and combining them together. The fundamental idea behind XGBoost is to iteratively build decision trees while giving more weight to the instances that were previously misclassified, thereby focusing on the areas where the model needs improvement.

XGBoost employs a variety of techniques to optimize performance and control overfitting, including regularization, gradient boosting, and pruning. It also incorporates advanced features such as handling missing values and providing flexibility in defining custom optimization objectives. XGBoost has proven to be exceptionally effective in various machine learning competitions and real-world applications due to its robustness and high predictive power.

2.2.4 Deep Learning and Pre-trained Models

Many modern advanced approaches no longer require manually designed features. Instead, they use self-supervised machine learning, where features are automatically generated by other machine learning models based on the context of the data. In recent years, with the rapid development of deep learning, there are many new applications to healthcare. Initially, researchers used feedforward neural networks (FNN) to classify healthy individuals and cognitively impaired patients (Rumelhart et al., 1986). For classifying speech or text data of Alzheimer's disease (AD) patients, which is a similar task with this project, researchers used CNN models and linear gated convolution neural networks (GCNN) (Najibi et al., 2016a). To capture timing information from speaker audio, they incorporated recurrent neural network (RNN) (Najibi et al., 2016b), long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997a), gated recurrent unit (GRU) (Chung et al., 2014), bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997), etc. Attention mechanisms (Vaswani et al., 2017a) also gained popularity, improving model accuracy by adding them to RNN, CNN, and LSTM models.

Deep learning methods have proven effective in the detection of AD, outperforming traditional machine learning approaches (Yang et al., 2022). This success has motivated us to explore the application of advanced deep learning models in the detection of TLOC causes, which holds the promise of achieving even better performance. To detect Alzheimer's disease with limited data, researchers use pre-trained models like Longformer, BERT, and ERNIE, which have been trained on large datasets, to help extract more effective features. For instance, BERT can create numerical vectors for words, ensuring words with similar meanings tend to have numerical representations that are close or similar to each other. The experiment compared methods with and without pre-training, and the results indicated that using pre-training is more beneficial than training models from scratch (Figure 2.2) (Yang et al., 2022).

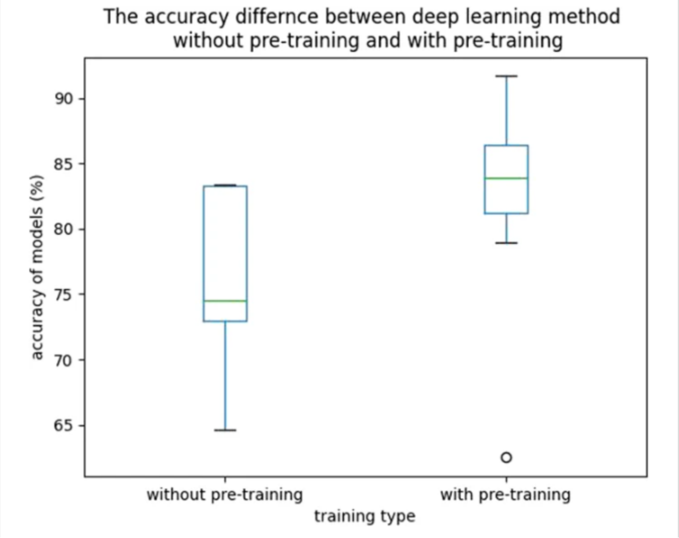


Figure 2.2: Comparison of deep learning methods without and with pre-training (Yang et al., 2022)

Methods for the ADReSS Challenge

The ADReSS Challenge is a recent worldwide competition focused on using speech to detect Alzheimer's disease. Among the top five teams in the competition, many of them used pre-training methods and two powerful deep learning techniques (Yang et al., 2022). The initial method includes training a deep learning structure on extensive datasets first and then adjusting it further using the ADReSS dataset. BERT (Saltz et al., 2021) model for pre-training, achieving ADReSS test set accuracies of 90%. The second way is extracting features from deep learning architectures, and then using these extracted features to train traditional machine learning classifiers. One approach is to blend traditional language characteristics with linguistic embeddings taken from a pre-trained BERT-based model (Syed et al., 2021). They honed these features through ensemble learning and combined them using majority voting, which resulted in an accuracy of 91.67% on the ADReSS test set.

When dealing with limited clinical data, it's crucial and effective to select appropriate pre-trained tasks and fine-tune models for disease classification. A commonly used method entails first pre-training a speech or text encoder on a large dataset and then utilizing attention mechanisms to establish connections within the data. Following this, a fine-tuned model is trained on AD or MCI datasets to create a framework for Alzheimer's disease classification from the ground up.(Yang et al., 2022).

Transformer-based Model for Detecting Schizophrenia

A new model based on transformers has been proposed to automatically evaluate the severity of thought disorder in schizophrenia (Huang et al., 2022). This model breaks the prediction process into three parts, which could potentially lead to better outcomes. For Semantic Representation, BERT (Devlin et al., 2018) is used as the text representation method to understand context of the text. Unlike Word2Vec (Mikolov et al., 2013), BERT pays attention to the words around it using self-attention.

For sentence structure information, a pre-trained model is used to extract features. ELECTRA's pre-training task, called "replaced token detection," helps the model figure out whether tokens in a sequence have been swapped or not. This model achieved a high prediction accuracy of 88%.

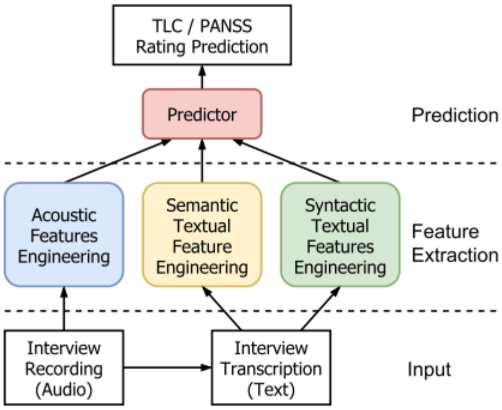

TLC / PANSS
Rating Prediction
Predictor
Prediction
Acoustic Features Engineering
Semantic Textual Feature Engineering
Syntactic Textual Features Engineering
Feature Extraction
Interview Recording (Audio)
Interview Transcription (Text)
Input

Figure 2.3: System overview of the schizophrenia prediction model (Huang et al., 2022).

2.3 Automatic Speech Recognition (ASR)

2.3.1 The Application of ASR in TLOC Diagnosis

In certain cases, a machine learning algorithm may require a transcript of the speech recording, especially when the features used to train the algorithm involve semantic content. In the process of assessing the performance of the conversation analysis model, both manually created and automatically generated transcriptions can be used.

While manually processed transcripts might boast a minimal word error rate and furnish invaluable insights into well-scrutinized research data, effectively implementing the TLOC assistance diagnosis system in real-world scenarios necessitates the integration of ASR system. Achieving nearly error-free text through manual transcription is feasible, yet in certain medical applications, the intricate and imbalanced nature of data perpetuates a substantial gap between machine-generated transcriptions and their human counterparts. Enhancing the precision of ASR systems, fortifying their resistance to interference, or bolstering their transcription performance under challenging conditions can substantially mitigate the word error rate in transcribed text.

The quality of transcribed text directly affects the efficacy of feature extraction. A diminished word error rate corresponds to more meticulous and uniform extracted features that seamlessly encapsulate patient information. As a result, this leads to increased accuracy when formulating diagnostic predictions, a critical consideration for both patients and their families.

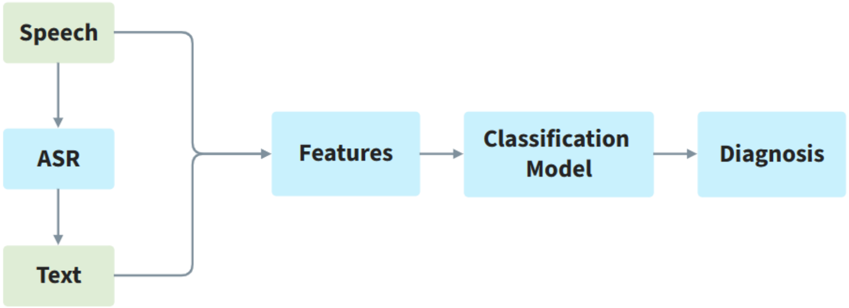


Figure 2.4: Block diagram of the automatic TLOC diagnosis system.

Incorporating an ASR system into the diagnostic prediction process can automate the diagnostic process, but the cost of the process is a reduction in accuracy. There was research that implemented ASR in a TLOC diagnosis machine learning system (Pevy, 2023). The Word Error Rate (WER) of this ASR module is 39.28%. As we discussed earlier, incorporating ASR negatively affected the performance of language analysis. The accuracy of distinguishing between epilepsy and Functional Dizziness Syndrome (FDS) using binary classification dropped by 24.5% when considering formulation effort features and LIWC (Linguistic Inquiry and Word Count) semantic categories, and by 12.2% when using TFIDF (Term Frequency-Inverse

Document Frequency) features. In a dementia diagnosis system (Mirheidari et al., 2016), the noisy ASR transcripts result in a decrease in the effectiveness of semantic features from 85% to 67%. Similarly, the accuracy of acoustic features drops from 77% to 67% . From these findings, it is evident that the transcription quality of the ASR system has a significant impact on the diagnostic system. To improve the accuracy of TLOC diagnosis predictions, more advanced ASR models are necessary.

2.3.2 Mel-Frequency Cepstral Coefficients (MFCC)

The waveform in the time domain is the record of a speech signal information. In order to convert speech signal into a series of statistical information, signal should be converted from each segment into a restricted set of features (Alim and Rashid, 2018), MFCC is a widely adopted technique for extracting these essential features from speech signals. It acts as feature extraction in the ASR process, as shown in Figure 2.5.

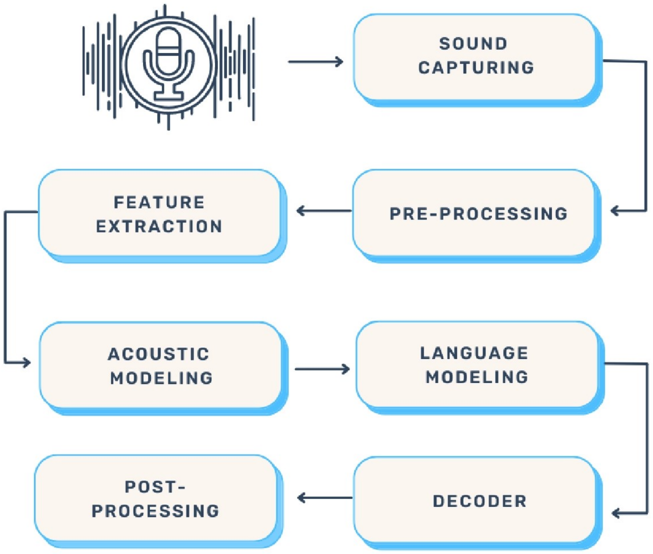


Figure 2.5: Feature extraction in traditional ASR Structure (Karatas, 2023)

These coefficients capture critical spectral characteristics and are well-suited for audio analysis tasks. The Figure 2.6 gives an overview of the MFCC process.(Basak et al., 2023):

- Analog-to-Digital Conversion: Initially, sound exists in an analog format as a mechanical wave. To use audio in computers and machine learning systems, it must be converted into digital signals. This is achieved through an Analog-to-Digital Converter (ADC) with a suitable sampling frequency based on the application.

- Pre-emphasis: Raw speech signals may suffer from rapid energy reduction, causing implementation challenges. To tackle this, a pre-processing technique called pre-emphasis

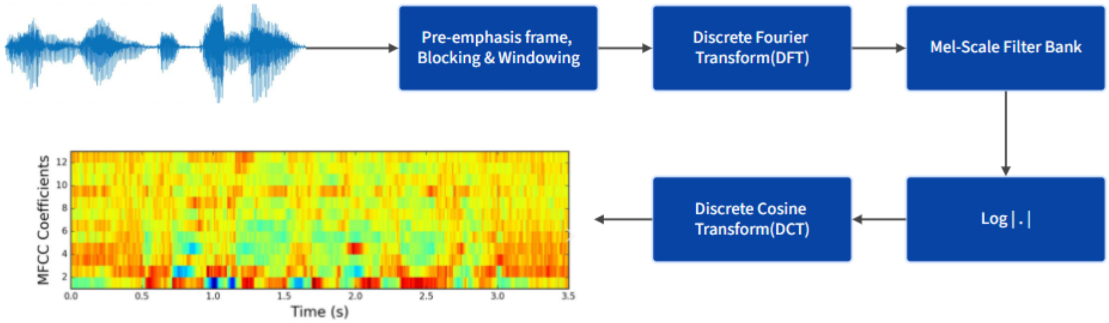

Pre-emphasis frame, Blocking & Windowing
Discrete Fourier Transform(DFT)
Mel-Scale Filter Bank
Log | . |
Discrete Cosine Transform(DCT)
MFCC Coefficients
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Time (s)

Figure 2.6: Overview of the MFCC process (Nair, 2018)

is applied. It emphasizes higher frequencies and compensates for the average spectral shape using a first-order Finite Impulse Response (FIR) filter.

- Framing: The speech signal is segmented into frames, each containing a specific number of samples. This process is known as framing. Dividing the signal into short frames allows us to assume that the speech signal is stationary within each frame, simplifying Fourier analysis.

- Windowing: The framed signal is subjected to windowing to smooth the endpoints, eliminating discontinuities and enhancing harmonic sharpness. The commonly used window types are Hamming and Rectangular windows.

- DFT (Discrete Fourier Transform): The Discrete Fourier Transform is a method used to change audio signals from the time domain to the frequency domain. This helps in analyzing speech signals more easily and offers valuable information for speech analysis. Ultimately, it results in the generation of the power spectrum for each frame of the signal.

- Mel Filter Bank: The signal is further processed using a Mel filter bank, which emulates the human ear's perception of sound by distinguishing sounds at lower frequency ranges. It maps the actual frequency to the frequency perceived by humans using the Mel scale.

- Applying Log: The output of the Mel filter bank undergoes a logarithmic function to simulate the human ear's response. This step is crucial as it produces a high gradient for low input values and vice versa.

- DCT (Discrete Cosine Transform): The resulting spectrogram from the Mel filter bank is decorrelated using the Discrete Cosine Transform (DCT). This transformation converts the Mel spectrum back into the time domain, resulting in a list of integers known as Mel Frequency Cepstral Coefficients (MFCCs).

By using MFCCs, we can represent speech in a manner that is both informative and robust to variations in pronunciation, intonation, and background noise. This transformation allows us to extract the relevant statistical information needed for a wide range of applications in the field of speech and audio processing.

2.3.3 Acoustic Model and Language Model

The backend of an ASR system, shown in Figure 2.5, includes acoustic model and language model. The acoustic model listens to audio and figures out the chances of each character in the alphabet being spoken. Then, the language model steps in to turn these chances into words that make sense in a sentence. This language model, sometimes called the scorer, assigns probabilities to words and phrases based on what it learned from its training data. These models are usually created using a variety of machine learning and deep learning techniques (Dua et al., 2023).

Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a widely used machine learning technique, especially in acoustic modeling. It deals with sequences of hidden states that cannot be directly observed, hence the name ”Hidden” Markov model. The main goal of HMM is to uncover these hidden states and their associated probabilistic functions, even when the source or generating states for observations are unknown. Essentially, HMM represents a statistical model where the system is treated as a Markov process with uncertain variables. While the states themselves are unknown, the variables controlled by these states are known. Consequently, each state is linked to a probable classification for potential output tokens. The sequence of states is inferred from the sequence of tokens generated by the HMM (Wang et al., 2019).

HMM builds probabilistic models using common terms and assesses how likely each model is to produce an unknown utterance. It arranges feature vectors into a Markov matrix (chains) that includes probabilities for transitioning between states (**?**). By considering code words as states, HMM divides the signal's feature vector into different states and computes the probabilities of moving from one state to another (Chavan and Sable, 2013).

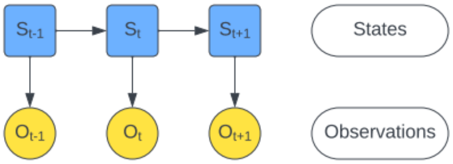


Figure 2.7: The HMM Module

Deep Learning Approaches

Deep learning is great for speech recognition because it can handle big, complex datasets well. It's good at working with the raw audio signals and feature data that we get from different techniques. In the backend of a speech recognition system, we can use various deep learning models, either alone or together. The hidden layers in these networks can extract useful information directly from the raw audio data when we input it into the model (Dua et al., 2023).

- Recurrent Neural Networks (RNN)
  RNN is effective in capturing the temporal evolution of voice signals. RNN's ability to remember a sequence of complex past events is achieved through a cell mechanism, which allows it to generate an output sequence at each time step based on the input vector it receives.

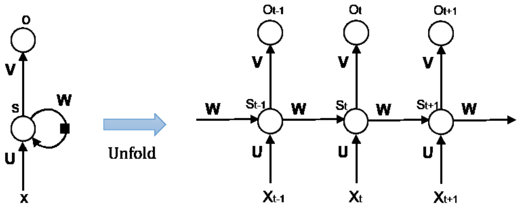


Figure 2.8: The RNN module (folded (left) and unfolded (right))

- Long Short-Term Memory Network (LSTM)
  LSTM (Hochreiter and Schmidhuber, 1997b) addresses the issue of long-term dependence, distinguishing it from a regular RNN, as it can retain information for extended periods. The LSTM structure comprises four interconnected neurons, distinct from the repetitive nature of the RNN. LSTMs overcome the gradient vanishing problem associated with CNNs and offer improved accuracy. Each LSTM block consists of memory units (Figure 2.9) governed by three multiplicative gates: the forget gate, input gate, and output gate. These gates open and close based on exposure to memory blocks, allowing cells to gather information until the forget gate is triggered.

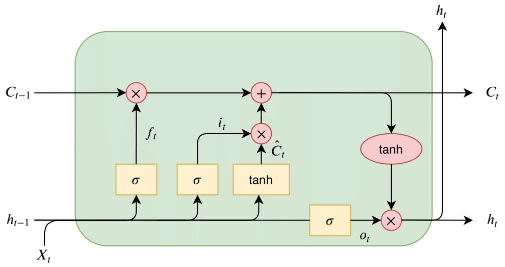


Figure 2.9: A Long Short-Term Memory Unit

- Transformer-based Model

  The self-attention mechanism plays a central role in the Transformer model (Vaswani et al., 2017b). In this autoregressive generative model, each layer has two parts: a self-attention part and a feed-forward part. What makes the Transformer different is that it doesn't use the usual RNN structure in both the encoder and decoder. Instead, it relies on attention mechanisms throughout the model, completely replacing the traditional RNN setup. (Figure 2.10).

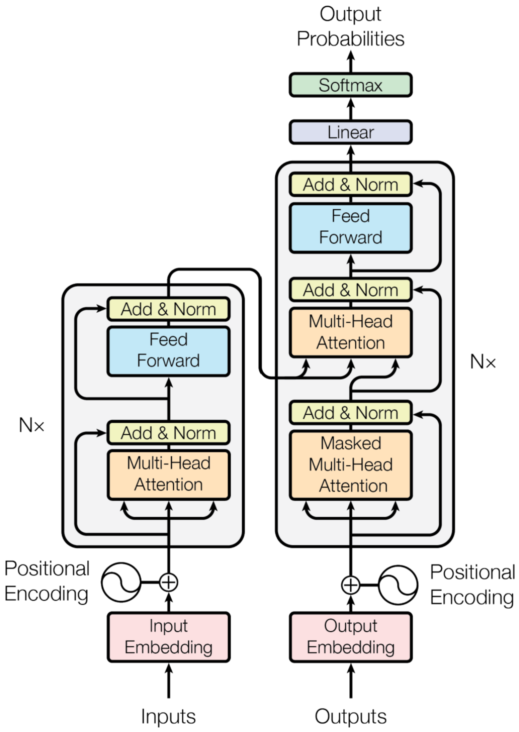


Figure 2.10: The Transformer Architecture (Vaswani et al., 2017b)

The Self-Attention feature enables simultaneous attention to all tokens within a group, allowing for parallelised computations and significantly reduced processing time. In this mechanism, tokens in the input sequence are transformed into vectors representing queries, keys, and values. Attention weights are computed using queries and their corresponding keys. Multiplying the values by these attention weights yields the output of the attention layer, as demonstrated in formula 2.3.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2.1}$$

$$\text{where} \quad d_k = \text{dimension of keys}$$

The transformer-based acoustic models (AMs) (Wang et al., 2020) demonstrate remarkable word error rate (WER) improvements compared to powerful bi-directional LSTM (BLSTM) baselines. These improvements were observed on both the widely-used Librispeech benchmark.

2.3.4 Pre-trained End-to-End ASR Models

The End-to-end (E2E) approach unifies the separate acoustic and language models from the traditional method and integrates them into one system.

ESPnet

ESPnet (Watanabe et al., 2018) fully utilises the benefits of two major end-to-end ASR implementations based on both connectionist temporal classification (CTC) and attention-based encoder-decoder networks (Figure 2.11).

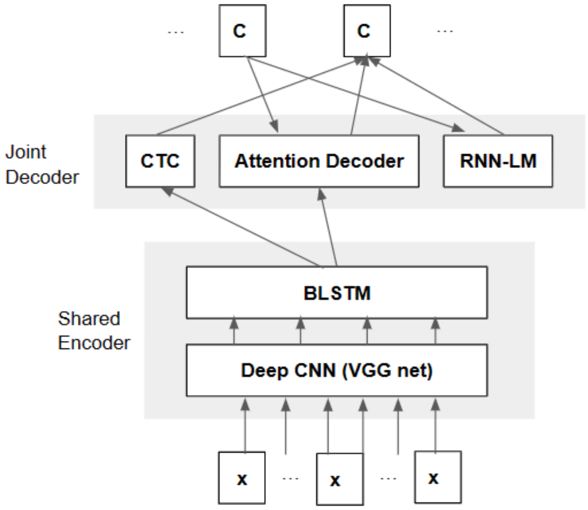


Figure 2.11: The ESPnet Architecture (Watanabe et al., 2018)

Attention-based methods use a special mechanism to align acoustic frames and recognized symbols, while CTC (Connectionist Temporal Classification) efficiently solves sequential problems using Markov assumptions. ESPnet, in its end-to-end ASR (Automatic Speech Recognition) approach, combines both of these methods for better results in training and decoding (Watanabe et al., 2018).

In training, ESPnet uses a multiobjective learning framework to improve accuracy when the alignment between audio frames and symbols is not perfect, and it helps the model learn quickly.

In decoding, ESPnet uses a one-pass beam search algorithm that combines both attention-based and CTC scores to find the best transcription, reducing errors caused by imperfect alignment between audio and symbols. This combination of techniques makes ESPnet more robust and accurate in speech recognition. (Watanabe et al., 2018).

Whisper

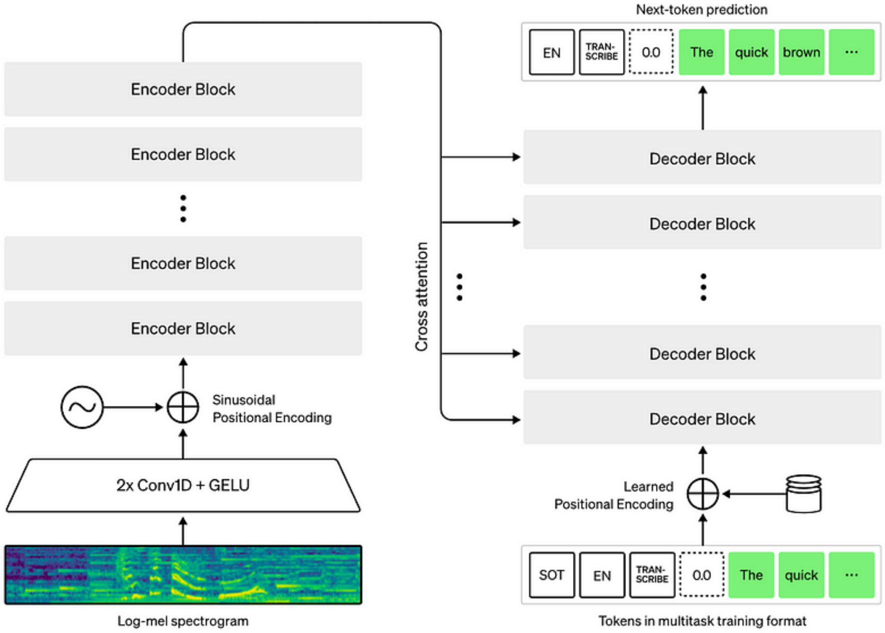


Figure 2.12: The Whisper Architecture (Radford et al., 2022)

Whisper models, trained on extensive and diverse audio datasets and evaluated in zero-shot scenarios, hold the potential to closely mimic human behaviour, outperforming existing systems.

In contrast to those larger models that boast intricate architectures and require expansive datasets for training or fine-tuning model parameters (Bert, GPT-3, VGG), Whisper places a stronger emphasis on model generalisation. And its architecture is not as complex as ESPnet (Figure 2.12). This is attributed to its exposure to expansive and varied datasets during training, which empowers it to demonstrate remarkable performance even within domains with limited data resources. This holds particular significance for applications in medical domains where acquiring substantial patient data remains challenging.

Functioning as a robust standalone speech processing system, Whisper delivers reliable performance without necessitating specific fine-tuning on individual datasets, ensuring high-quality outcomes across distinct distributions.

Whisper expands the realm of weakly supervised speech recognition significantly, leveraging a staggering 680,000 hours of labelled audio data. It demonstrates seamless applicability to existing datasets without any dataset-specific fine-tuning, yielding superior results (Radford et al., 2022).

2.4 Summary

In this chapter, we have reviewed some approaches based on clinical conditions used to detect the cause of TLOC. We explored the differences and connections among the three main causes. Most of these approaches rely on clinical characteristics, which require careful recording and significant effort. However, there have been some advancements in machine learning methods, such as iPEP, which can comprehensively consider patients' medical history and behaviours.

Inspired by these clinical decision tools, features extracted from patients' conversations are now being considered as important evidence or clues that indicate the possible cause of TLOC. As a result, there has been an increasing trend of applying conversation analysis in epilepsy classification, utilising machine learning models.

To achieve automatic diagnosis prediction, the ASR system is necessary to generate accurate transcripts of the conversation recordings. We explained the process of ASR in this chapter. Additionally, common models in ASR, such as HMM, RNN, LSTM, Transformer-based models and the state-of-the-art pre-trained E2E ASR model, were introduced in this chapter.

Chapter 3

Methodology

This chapter will introduce the details of the dataset in this project and data preparation. Then, we will explain the method to explore whether formulation effort features can be used to differentiate people with one of three main TLOC causes. Also, we will analyse the feasibility of extracting semantic features using the Linguistic Inquiry and Word Count (LIWC) approach. The classic machine learning models used to predict diagnosis will be mentioned in this chapter. Following that, we will discuss the weighting and threshold adjustment methods that can significantly improve the model performance. Additionally, we will explain the architecture of ASR models used in this project - Whisper and ESPnet. Last but not least, we will provide details of the metrics applied to the diagnosis prediction task and the ASR task. In the next chapter, we will report the various experiments identified in Chapter 2, and their results will be discussed in detail.

3.1 Dataset

The data for this project comes from past research on conversational analysis conducted at the Royal Hallamshire Hospital in Sheffield. All of these recordings and transcripts have been employed in previous TLOC research (Pevy, 2023). The people included in the earlier TLOC studies were those currently being examined at the Royal Hallamshire Hospital to figure out why they were having seizures. These participants gave their permission for their data to be used. The actual diagnosis for these patients was determined later on through clinical assessment or the video-EEG method. When they took part in the study, they hadn't received a final diagnosis yet. This dataset contains a total of 504 recordings from 53 patients.

Table 3.1: Percentage distribution of recordings

Diagnosis	**Number**	**Percentage**
Epilepsy	140	27.78%
FDS	265	52.58%
Syncope	99	19.64%

Among these patients, the percentages for epilepsy, FDS, and syncope are 32%, 47.1%, and 27.7%, respectively (Table 3.1). The patient information summary comprises 74 variables, including personal information, medical histories, diagnosis, and feedback. Given that the majority of the information is unrelated to voice-related data, only two columns, 'labels' and 'patient numbers,' will be extracted from the data summary.

3.2 Data Preparation

The audio files were split into 30-second parts, and each part was matched with the corresponding piece of the transcript for that time. We used all these audio segments for training, even if they didn't contain speech, but we did so with a certain probability of sub-sampling. The text files contained manual transcripts of the patients' responses to questions. We processed the text by converting it to uppercase, removing punctuation and numbers, expanding contractions, and simplifying words to their base forms using the Natural Language Toolkit (NLTK) (Loper and Bird, 2002).

3.3 Feature Extraction

3.3.1 Formulation Effort

Numerous acoustic attributes can be quantified by employing readily available software. In this experiment, we employ the Librosa package for tasks such as audio file loading and energy computation. Typically, the desired characteristics are computed for individual segments of an audio file, and we subsequently derive descriptive statistics to portray each attribute across the entire recording. Seven features shown in Table 3.2 were chosen as features of the formulation effort.

Among these features, three of them involved searching through the whole transcript for given words or keywords (Pevy, 2023). These features are the total number of hesitations (e.g. hm,erm or um), the total number of repetitions (e.g. I was leaving leaving home). One dictionary was created to record the appearance frequency of the keywords. The presence or absence of words that suggest uncertainty (e.g. sort of or maybe). A list of keywords associated with uncertainty was prepared. If any word in the uncertain word list appears in the transcript file, return 1; Otherwise, return 0. The total number of words and the speaking duration used to describe the onset experience and scene have also been proven can be used to help differentiation. So the word count method is to write a word count function.

The other three features involved measuring pauses within the interaction. A voice activity detector (VAD) specially used to detect non-voice parts was implemented using the Librosa library. Firstly, short-term energy is calculated using the root mean square (RMS) of the audio signal. This is the way to measure the energy in the audio signal over each 10 millisecond frame. Then, two thresholds are designed to detect pause activities. We set a pause threshold to identify pauses that are longer than 30 milliseconds. The next step is to iterate through the energy values, the energy value that falls below 0.1 will be considered as

Features	Definitions
Number of hesitations	The frequency of hesitations within the patient's speech based on a pre-specified list of possible hesitations (”HM”, ”UM”, ”ERM”).
Number of repetitions	The frequency at which a word (N) is a repeat of the previous word (N-1) or the word before (N-2).
Presence or absence of keywords associated with uncertainty	This feature indicates whether any of a list of keywords associated with uncertainty appears in the seizure description or not.
Pause frequency	The percentage of pauses that were greater than 30 ms in length in the entire patient's recording file.
Average pause length	The average length of all patient pauses that were greater than 30 ms.
Total pause time	The total time spent pausing when pauses were defined as being longer than 30 ms.
Word count	Total number of words in the transcripts.
Speech duration	Duration of a patient speaking in the audio file.

Table 3.2: Features and definitions (Pevy, 2023)

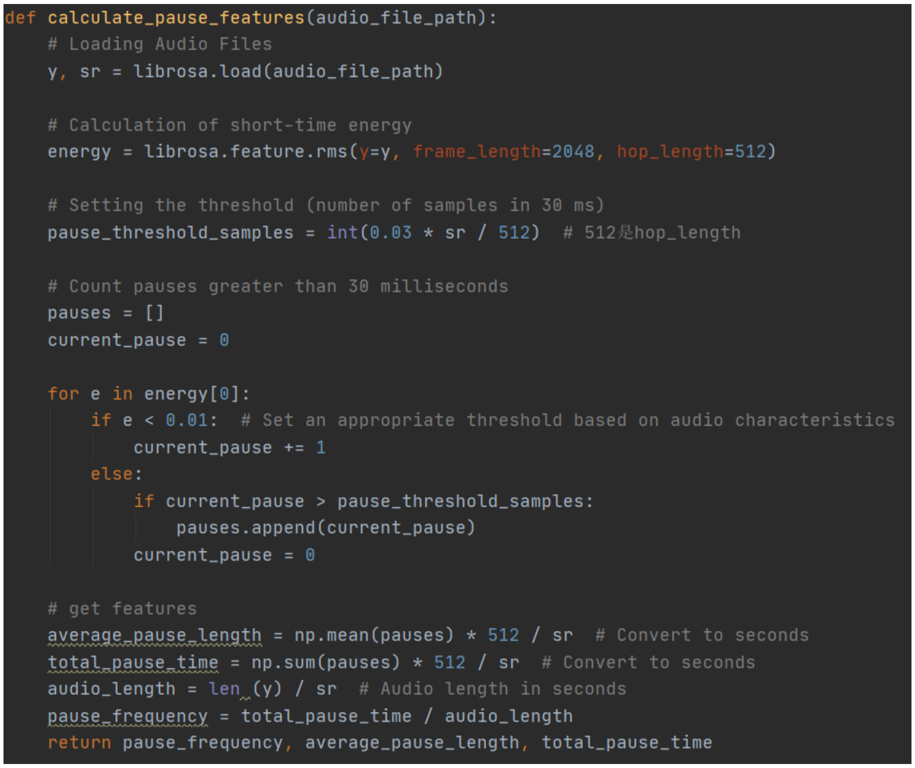
a candidate for a pause. We keep track of the number of consecutive frames with low energy, and it will be counted as a pause if the consecutive frames keep staying lower than the energy value threshold within 30 ms. After detecting and recording the pauses, three pause-related features are defined as average_pause_length, total_pause_time, and pause_freaquency.

3.3.2 Semantic Categories

Linguistic Inquiry and Word Count (LIWC) is a software application designed for text analysis. It evaluates the proportion of words that fall into various semantic categories (Pennebaker and Graybeal, 2001).

This tool provides a specialised method for comprehending, interpreting, and quantifying psychological, social, and behavioural phenomena. By looking at how people use words, we can get clues about how they feel, what they're thinking, and what's on their minds socially. For example, if someone often says words like ”happy,” ”excited,” and ”joyful,” it probably means they're feeling happy right now. A tool called LIWC checks the text you give it, compares each word with a list of words it knows, and figures out the percentage of words in different categories. For example, if you put in a 1,000-word speech and LIWC's dictionary has 50 positive emotion words and 10 belonging words, it might tell you that 5.0% of the words are about positive emotions and 1.0% are about belonging. (Mehl, 2006).

These linguistic characteristics are centred around the language patterns exhibited by individuals with specific health conditions. For instance, they involve gauging the prevalence



```python
def calculate_pause_features(audio_file_path):
    # Loading Audio Files
    y, sr = librosa.load(audio_file_path)

    # Calculation of short-time energy
    energy = librosa.feature.rms(y=y, frame_length=2048, hop_length=512)

    # Setting the threshold (number of samples in 30 ms)
    pause_threshold_samples = int(0.03 * sr / 512)  # 512是hop_length

    # Count pauses greater than 30 milliseconds
    pauses = []
    current_pause = 0

    for e in energy[0]:
        if e < 0.01:  # Set an appropriate threshold based on audio characteristics
            current_pause += 1
        else:
            if current_pause > pause_threshold_samples:
                pauses.append(current_pause)
            current_pause = 0

    # get features
    average_pause_length = np.mean(pauses) * 512 / sr  # Convert to seconds
    total_pause_time = np.sum(pauses) * 512 / sr  # Convert to seconds
    audio_length = len (y) / sr  # Audio length in seconds
    pause_frequency = total_pause_time / audio_length
    return pause_frequency, average_pause_length, total_pause_time
```

Figure 3.1: Code segment for pause detecting

of keywords related to various semantic categories, assessing speech tempo, evaluating features aimed at quantifying linguistic intricacy, and analysing the frequency of specific parts-of-speech labels (Matton et al., 2019a). The doctor-patient interactions' audio recordings underwent manual transcription, followed by the utilisation of a custom algorithm to extract text corresponding to all patient responses during the designated portion of the interaction.

We opted to employ 29 specific semantic categories (Table 3.3), tailored to measure variances observed in prior research and explore potential associations among these independent categories. These categories are associated with the six exhibited characteristics.

1. When people with FDS experienced a seizure with others present, they exhibited different behaviours compared to individuals with epilepsy. This included more efforts to alert others about an impending seizure, heightened intensity, and divergent seizure-after behaviour. Earlier investigations also indicated discrepancies in language usage between these groups. People with epilepsy tended to employ words related to categories such as ”We,” ”She/He,” and ”Family Reference” (Cardea et al., 2020)

2. To assess tendencies towards catastrophizing, the category ”Risk” was included (Whitfield et al., 2020), alongside ”Cause” to detect inferences regarding the consequences of seizures on daily life, and ”Reward” as a countermeasure.

3. Furthermore, thematic analyses of written accounts of epilepsy and FDS revealed

Number	Definitions and Categories
1	The attempts to using third party references Categories: We, She/He, Family, Social, Affiliation, Friend
2	The consequence of seizures on everyday life Categories: Risk, Cause, Reward
3	Emotional experience and expression Categories: Emotion, Affect, Positive Tone, Negative Tone, Positive Emotion, Negative Emotion, Tone, Affect, Anxiety, Anger, Sad
4	Describing a single experience that happened in the past or multiple seizure experiences in the present tense Categories: Memory, Focus Present, Quantity, Certitude, Number
5	Differences in tentative language associated with formulation effort Categories: Tenant, Discrep
6	The way patients conceptualize the seizure Categories: Space, Power

Table 3.3: Definitions and categories (Pevy, 2023)

disparities in the ”emotional tone” of these accounts (Pevy, 2023). Those with epilepsy typically conveyed relatively stable moods, whereas individuals with FDS exhibited greater anxiety and lower mood (Rawlings et al., 2017a). Hence, seven categories were integrated to gauge variations in emotional language, encompassing ”Emotional Tone,” ”Affect,” ”Positive Emotions,” ”Negative Emotions,” ”Anxiety,” ”Anger,” and ”Sadness” (Rawlings et al., 2017b).

4. Additionally, we included ”Focus Present” and ”Quantifiers” categories to ascertain whether individuals described single past experiences or multiple present-tense seizure episodes. For instance, distinguishing between ”I never remember what happened in the seizures” and ”I sometimes lose consciousness” as opposed to ”I was walking down the street and began to feel strange.” Moreover, the ”Certainty” category was utilised to capture holistic statements and complete negations, such as ”I never remember anything.” (Pevy, 2023)

5. It was noted that individuals with epilepsy often employed metaphors portraying seizures as agents or forces that could be combated, whereas those with FDS more frequently used metaphors conceptualizing seizures as spaces or places they entered (Plug and Reuber, 2009). To capture some of these distinctions, we introduced the categories ”Space” and ”Power.”

6. The manner in which patients perceive seizures, whether as an entity or a destination

they enter, can vary. Additionally, we utilised the semantic category ”Tentativeness” to detect distinctions in language hesitancy related to the effort of articulating experiences, such as an increasing use of tentative language when describing subjective symptoms.

3.3.3 Classification Models

Classical Machine Learning Models

We use several machine learning classifiers to determine the most suitable algorithms for this particular TLOC cause detection task. The semantic categories are the input of four machine learning models: Random Forest, Support Vector Machine, XGBOOST and K-Nearest Neighbor. Besides, the cross-validation is used in the training process to avoid the overfitting.

Threshold and Weight Balancing During Training

In this project, we introduced a threshold factor as a strategy to prevent the prediction results from being heavily skewed towards a single class and to enhance result balance. This factor serves to adjust the threshold values applied to the probability predictions for each class, ultimately transforming these predicted probabilities into definitive class labels. This process involves assigning samples with predicted probabilities exceeding the threshold as positive class instances, while those falling below the threshold are designated as negative class instances. Positive class instances receive a weight of 1, while the negative class instances remain unaffected. For instance, consider the XGBoost model employed in this project, which utilises a threshold of ’[0.3, 0.6, 0.2]’. If, for a particular sample, the model generates predictions of ’[0.1, 0.5, 0.5]’, only the probability for the third class exceeds its corresponding threshold. Consequently, the class prediction result transforms to ’[0.1, 0.5, 1.5]’, with the second class, having the highest probability, being designated as the predicted diagnosis. It’s important to note that the choice of thresholds can have a significant impact on prediction outcomes. By selecting appropriate thresholds, it is possible to optimise the model’s performance.

3.3.4 ASR Models

Whisper

Whisper is a general-purpose speech recognition model (Radford et al., 2022). The focus of Whisper is on studying the capabilities of large-scale supervised pre-training for speech recognition. The encoder-decoder Transformer is used in Whisper because this architecture has been well-validated to scale reliably. All audio is resampled to 16,000 Hz and a log-amplitude Mayer spectrogram is computed for 80 channels over a 25 ms window in 10 ms steps (Radford et al., 2022).

For feature normalisation, Whisper employs a global scaling method that controls the input values in the range of -1 to 1, while maintaining an approximate centred mean on the pre-training dataset. The input representation is then processed by an initial segment consisting of two convolutional layers, each with a filter width of 3, and using the GELU activation function (Hendrycks and Gimpel, 2016).

Following this, sinusoidal position embeddings are incorporated into the output of the initial segment. Subsequently, the encoder applies Transformer blocks. These Transformer blocks employ pre-activation residual architecture (Child et al., 2019), and a final layer normalisation is applied to the encoder's output. On the other hand, the decoder employs learned position embeddings and tied input-output token representations (Press and Wolf, 2017). It's worth noting that the encoder and decoder share the same dimensions and number of transformer blocks.

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	122	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 3.4: Architecture details of the Whisper model family (Radford et al., 2022)

Whisper employed a variety of models with different specifications, including variations in the number of layers, layer width, the count of heads, and the overall number of parameters. Subsequently, we applied three models with different sizes (tiny, small, medium) to the TLOC cause prediction task in our project. Further details and insights can be found in the upcoming chapter. During training, the models were optimised using the AdamW optimiser (Loshchilov and Hutter, 2019) and included gradient norm clipping (Pascanu et al., 2015). A linear learning rate decay was implemented, gradually reducing the learning rate to zero over the initial 2048 updates. The training was carried out with a batch size of 256 segments, and the models underwent 2^{20} updates, which roughly translates to between two and three complete passes through the dataset.

ESPnet

ESPnet applies hybrid CTC/attention end-to-end ASR (Watanabe et al., 2017), which combines the advantages of both architectures in training and decoding. During training, the ESPnet combines CTC L^{ctc} and attention-based cross entropy L^{att} to improve effectiveness, as follows:

$$\mathcal{L} = \alpha\mathcal{L}^{ctc} + (1-\alpha)\mathcal{L}^{att}$$

The training method uses a CTC objective function as an auxiliary task to train the attention model encoder.

There is a tuning parameter α to linearly interpolate both objective functions and usually set as $\alpha = 0.5$(equal contributions)

During decoding (Watanabe et al., 2018), ESPnet performs joint decoding by combing both attention-based and CTC scores in a one-pass beam search algorithm to further eliminate

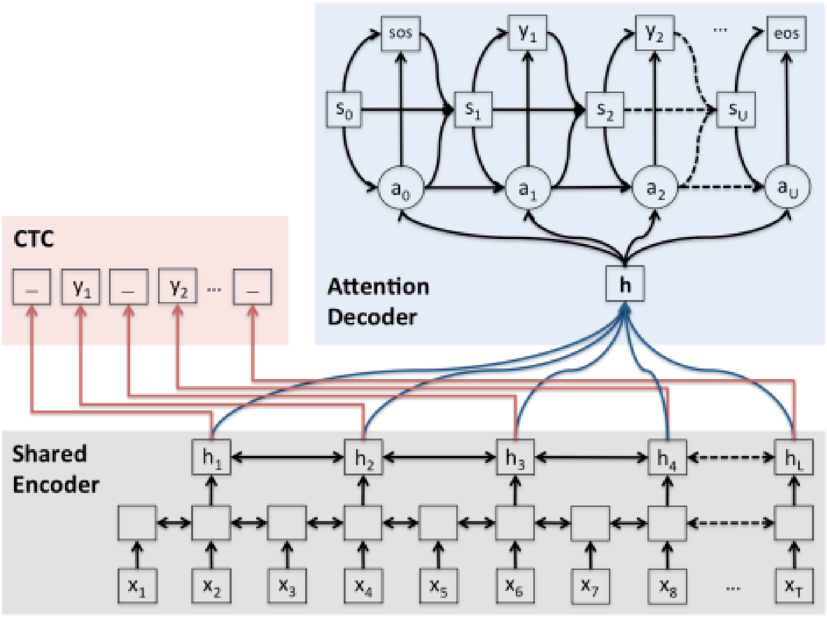


Figure 3.2: Joint CTC-Attention based end-to-end framework (Watanabe et al., 2018)

irregular alignments. This model takes the CTC probabilities into account to find a hypothesis that is better aligned to the input speech, as shown in Fig3.2.

Attention p_att and CTC p_ctc log probabilities are combined during the beam search (Watanabe et al., 2018):

$$\log p^{hyb}(y_n|y_{1:n-1}, h_{1:T}) = \alpha \log p^{ctc}(y_n|y_{1:n-1}, h_{1:T}) + (1 - \alpha) \log p^{att}(y_n|y_{1:n-1}, h_{1:T})$$

Implementation details of ESPnet

The standard recipe flow of ESPnet is shown in Figure 3.3.The Kaldi data preparation script `data_prep.sh` are used in data preparation step. Next, the 80 dimensional log Mel feature with the pitch feature will be extracted. Following that, all the vital information present within the Kaldi data directory is transformed.

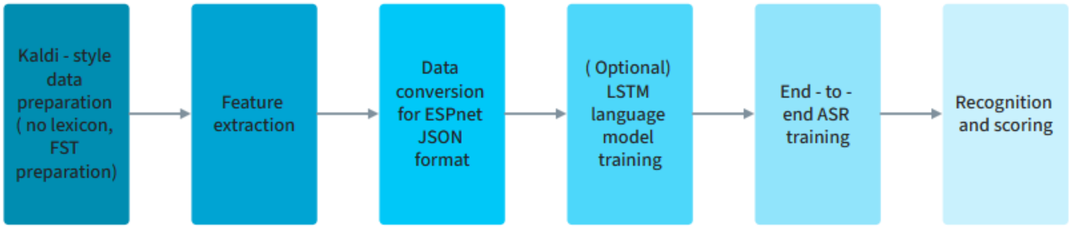

Kaldi - style data preparation (no lexicon, FST preparation)
Feature extraction
Data conversion for ESPnet JSON format
(Optional) LSTM language model training
End - to - end ASR training
Recognition and scoring

Figure 3.3: Experimental flow of standard ESPnet recipe (Radford et al., 2022)

The outcome is a transformed JSON file named `data.json`, excluding input features. Proceeding to Language Model Training, a character-based Recurrent Neural Network Language Model (RNNLM) is trained. Then a hybrid CTC/attention-based encoder-decoder model is trained. Lastly, the character-basd RNNLM and the end-to-end ASR model will be applied in speech recognition tasks.

3.3.5 Evaluation

Classification Performance Metrics

To assess the model's performance in this project, we employ several evaluation metrics. Accuracy gauges the proportion of accurate predictions made by the model. Precision measures the percentage of correct positive predictions for a particular class. Recall, also known as Sensitivity, assesses the model's ability to accurately recognize individuals belonging to a specific class.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

- **TP (True Positives)**: It represents the number of instances where the model correctly recognized a condition or class.
- **TN (True Negatives)**: It represents the number of instances where the model correctly recognised the absence of a condition or class.
- **FP (False Positives)**: It represents instances where the model mistakenly identified a condition or class when it wasn't present.
- **FN (False Negatives)**: It represents instances where the model missed identifying a condition or class that was actually there.

The dataset is imbalanced with the majority of the samples with a diagnosis of FDS, while syncope samples rate no more than 20 percent. Only using Accuracy or recall as the metric will cause all samples will be classified as FDS and still achieve more than 50% accuracy. In this project, considering the imbalance of data, the F1 score is selected as the main metric.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

ASR Performance Metrics

- Word Error Rate (WER): WER is a crucial metric for evaluating the performance of Automatic Speech Recognition (ASR) systems. It measures the level of word-level mismatch between the output text generated by the ASR system and the reference text.

  $$WER = \frac{S + D + I}{N} \times 100$$

  where S represents substitutions,
  D represents deletions, I represents insertions, C represents correct words, N represents the number of words in the reference (N=S+D+C)

- Word Information Lost (WIL): WIL quantifies the tendency of an ASR system to insert additional vocabulary during the transcription process.

  $$WIL = \frac{I}{N} \times 100$$

  Where: I represents the number of inserted words. N represents the total number of words in the reference text.

- Relative Information Lost (RIL):RIL quantifies the amount of information lost by an ASR system during the recognition process.

  $$RIL = \frac{D + I}{N} \times 100$$

  Where: D represents the number of deleted words. I represents the number of inserted words. N represents the total number of words in the reference text.

- Word Information Preservation (WIP):WIP evaluates whether an ASR system preserves valuable information from the input speech during transcription.

  $$WIP = 100 - RIL$$

Chapter 4

Experiment and Discussion

4.1 Experiment One

The baseline model is the application of classic machine learning models to make predictions on patients' data. The formulation effort features and semantic categories features are used to train various models respectively.

4.1.1 Baseline1

As we mentioned in the last chapter, eight features were designed as markers of the formulation effort: Number of hesitations, Number of Repetitions, Presence or absence of keywords associated with uncertainty, Pause Frequency, Average pause length, Total pause time, Word count, Speech duration.

Four machine learning algorithms, Random Forest, Support Vector Machine, XGBOOST and K-Nearest Neighbour, are used to investigate whether the formulation effort features were capable of differentiating TLOC causes.

The audio data was divided into training, validation, and test sets in an 80-10-10 ratio. The format of the extracted features is data list. It represents formulation effort from the patients' audio files and their handcrafted transcripts. These features serve as input for the models. We expanded the sample size by using individual questions as the speech units, rather than representing them in the form of individual patients. In other words, the data used in this experiment consists of 520 audio segments, each corresponding to answers to various seizure-related questions. Each segment is labelled with the respective disease category (epilepsy, FDS, or syncope). Accuracy performance of the models trained with formulation effort are shown in Table 4.1. It can be observed that out of the four models, three of them have an accuracy higher than 50%. The XGBOOST model performs the best on the test set. Delving into the factors behind this performance, a portion of it can be credited to the configuration of thresholds and weight balance within the XGBOOST model, rendering it a more fitting choice for the dataset in this project.

We performed an error analysis on the results, and this table displays the prediction

	Dev Accuracy	Average	Test Accuracy	Average
XGBOOST	0.469	0.4845	0.596	**0.545**
	0.484		0.576	
	0.485		0.455	
	0.5		0.553	
RF	0.471	0.49175	0.565	0.5245
	0.509		0.525	
	0.5		0.463	
	0.487		0.545	
SVM	0.535	**0.51625**	0.439	0.46075
	0.523		0.471	
	0.518		0.459	
	0.489		0.474	
KNN	0.495	0.50425	0.545	0.51375
	0.52		0.482	
	0.5		0.51	
	0.502		0.518	

Table 4.1: Model accuracy performance on dev and test dataset

distribution of the best-performing XGBOOST model for each label in the above experiments.

	Test			Dev		
Label/Prediction	Epilepsy	FDS	Syncope	Epilepsy	FDS	Syncope
Epilepsy	**46.00%**	38.00%	16.00%	**46.92%**	39.23%	13.85%
FDS	32.67%	**56.67%**	10.67%	40.17%	**43.59%**	16.24%
Syncope	47.27%	23.64%	**29.09%**	44.32%	28.41%	**27.27%**

Table 4.2: Distribution of results for each label (The correct classification proportion for each label is highlighted in bold)

It can be observed that all three categories have a higher proportion of correct classifications on the test dataset than their chance probabilities (Table 4.3). Therefore, the use of formulation effort and traditional machine learning models for TLOC can help in the diagnosis, and it is effective for all three of the most common causes, epilepsy, FDS and syncope.

In practice, there are many effective and stable methods other than speech analysis, such as iPEP and ECG, to distinguish between seizure and syncope, but the accuracy in the diagnosis of the two types of seizure has not been high. Therefore, based on the practical needs, after verifying the effectiveness of this method for threeway classification, another objective of this experiment is to explore the effectiveness of this method for the binary classification between epilepsy and FDS. In this process, I removed the samples labelled as ”syncope” to observe the classification performance exclusively for the other two types of

label	Sample Number	Percentage
Epilepsy	140	27.78%
FDS	265	52.58%
Syncope	99	19.64%

Table 4.3: Number and chance probability for each label

seizures. After excluding "syncope," there remained a total of 406 samples.

Same as last step, each model was trained and subjected to inference on the dataset four times with identical configurations. The results were then averaged, and the final performance of the four models on four metrics: Accuracy, Precision, Recall, and F1-scoreis shown in the Table 4.4.

Model	Accuracy	Precision	Recall	F1-Score
RF	0.59975	0.61775	0.54525	0.55775
XGBOOST	0.6295	0.61925	0.57075	0.578
SVM	**0.63875**	**0.6355**	**0.58175**	**0.59325**
KNN	0.59875	0.606	0.56825	0.57175

Table 4.4: Models performance on dataset without syncope (Best performing model is highlighted in bold)

As observed in the table, after excluding "syncope," the accuracy of classification significantly improved, with each model achieving an accuracy of around 60%. Considering the data's imbalance, several other metrics further illustrate the models' performance on this project's dataset. Notably, SVM emerges as the top-performing model in every metric (bold numbers indicate the best performance). Therefore, we conducted an error analysis on the SVM model, and the Table 4.5 depicts the prediction distribution for two types of labels.

	Dev		Test	
Label/Prediction	Epilepsy	FDS	Epilepsy	FDS
Epilepsy	53.60%	46.40%	74.67%	25.33%
FDS	45.80%	54.20%	40.77%	59.23%

Table 4.5: Error analysis(without syncope)

This model achieved classification accuracies exceeding 50% for both classes in the validation set. On the test set, it achieved a diagnostic accuracy of 74.67% for epilepsy and a classification accuracy of 59.23% for FDS, demonstrating commendable classification performance. This indicates that the model can provide valuable assistance in diagnosis.

4.1.2 Baseline2

The objective of the second baseline experiment was to explore whether semantic differences can assist the differential diagnosis. The purpose is to assess the frequency of different semantic content within conversations.

As we introduced in Chapter 2, there are significant differences in LIWC categories between people with epilepsy and FDS (Cardea et al., 2020). Six characters with a total 29 categories are chosen in this project. For semantic categories features, we use them to train four machine learning models like the implementation in Baseline 1. Each of these models training uses a cross-validation strategy, coupled with a nested search for optimal hyperparameters to mitigate the risk of overfitting (Vabalas et al., 2019).

In order to minimise the effects of randomness, each model was trained and evaluated four times with the same configurations. The model parameters are initialised after each experiment to ensure that the model parameters will not be affected by the previous experiments. The performance of each model using semantic categories features is shown in Table 4.6.

Model	F1-score	Recall	Accuracy
XGBOOST	0.471	0.489	0.56375
RF	0.4475	0.462	0.57075
SVM	**0.4825**	**0.49825**	**0.59225**
KNN	0.47325	0.48525	0.55875

Table 4.6: Models performance on the dataset without syncope (Best performing model is highlighted in bold)

From this table, we can find the model with the best performance on semantic features is SVM. Compared with the model performance using formulation efforts, this model did not get a better performance.

We selected the SVM model that performed the best in this experiment for error analysis. The distribution of each label is depicted in table 4.5. For this model, the diagnosis of epilepsy is more accurate, whereas the model's discriminative ability for FDS is relatively less pronounced. It's worth noting that, compared to models trained using formulation effort features, this model exhibits improved discriminative performance for syncope.

To further investigate the effectiveness of this model in distinguishing between epilepsy and FDS, we also conducted an operation to remove samples labelled as syncope. The model's performance in classifying these two seizures is shown in Table 4.7.

Through this table, we can observe a significant improvement in accuracy, with the XGBoost model achieving the highest accuracy at 74.13%. The F1-score also reaches a high of 68.63%, a level that previous models did not attain. Therefore, currently, this model exhibits outstanding performance in diagnosing seizures and has a high level of discrimination between the two different seizure types.

Another significant discovery is that, contrary to expectations, the model's performance deteriorates after feature standardisation. Several factors may contribute to this phenomenon.

	Accuracy	Precision	Recall	F1-Score
RF	0.689	0.66375	0.6295	0.63625
XGBOOST	**0.74125**	**0.76025**	**0.6645**	**0.68625**
SVM	0.74	0.713	0.644	0.66075
KNN	0.66475	0.6505	0.628	0.63375

Table 4.7: Models performance on dataset without sync (Best performing model is highlighted in bold)

Firstly, standardisation assumes that features either follow a normal distribution or have similar scales. If these assumptions are not met by the data, standardisation may prove inappropriate. Secondly, standardisation transforms data to have a mean of 0 and a standard deviation of 1, which, in certain cases, can result in information loss, particularly when dealing with highly skewed data or data containing significant outliers. Additionally, there is the possibility that standardisation might make the model more susceptible to overfitting, especially when working with small datasets, as scaling can amplify the influence of outliers, potentially causing the model to overfit noise within the data.

4.2 Experiment Two

This experiment includes the implementation of two methods, Tokenizer class from Keras and Bert-based embedding, for text vectorization based on the LSTM neural network and the evaluation of these models on the patient data.

Since patients' descriptions of their medical conditions or seizure scenarios are essentially continuous narratives, the contextual relationship between preceding and subsequent text is a vital form of information. Properly leveraging the contextual information in sequential data might enhance prediction accuracy. Therefore, we also attempted to use LSTM (Long Short-Term Memory) models, as methods based on LSTM are variants of recurrent neural networks (RNNs). LSTM is specifically designed for handling sequential data, taking into account the sequential structure of sentences. In this experiment, we aimed to verify whether LSTM could better capture the long-term dependencies inherent in data such as patient recordings and transcripts, which possess extended contextual relationships. This, in turn, would facilitate a deeper understanding and processing of the data, ultimately leading to improved predictive performance.

4.2.1 Tokenizer from Keras Library + LSTM

The first method involves the use of Keras' Tokenizer class. Tokenization consists of dividing each sentence into individual words (tokens) and assigning a unique index to each word. This process is accomplished using the Tokenizer class from Keras.

The words are replaced with their corresponding indices, and the tokenizer also assigns the index 1 to the special token <OOV> (Out of Vocabulary), which will be used for words not present in the tokenizer's vocabulary. Afterward, sequences are padded with zeros to ensure they all have the same length. I set the padding length parameter, max_sequence_length, to 200. This conversion transforms words into padded indices, which is crucial for inputting the data into an LSTM neural network model.

However, this method resulted in overfitting, with all prediction outcomes concentrated in a single category, "epilepsy." I believe this is due to the model exceeding the adjustable complexity limit of this dataset, leading to classifier bias.

4.2.2 Bert-based Embedding + LSTM

BERT has the capability to provide rich contextual information, aiding in capturing sentence semantics more effectively. LSTM, on the other hand, can further process sequential information, particularly in tasks that require modelling contextual relationships within sequences. The implementation of this method is as follows:

- Pretraining BERT Model: Initially, a pre-trained BERT model trained on a vast corpus of text data is employed. During pretraining, BERT learns contextual relationships in language, generating rich word representations.
- Obtaining BERT Representations: For input text sentences, pre-trained BERT models are utilized to obtain contextually relevant representations for each word. BERT generates multi-layer representations for each word.
- Sequence Padding and Normalization: Input text sequences undergo sequence padding and normalization to ensure uniform sentence length.
- Input to LSTM: BERT representations serve as input to an LSTM model. At each time step, a word's BERT representation is input. The LSTM network further learns patterns and relationships within the sequence to adapt to specific tasks such as classification and labelling.

	Dev accuracy	Test accuracy
5 epoches	0.58	0.549
10 epoches	0.6	0.5294

Table 4.8: Accuracy performance of Bert-based embedding with LSTM

Following that, performance evaluation was carried out using a model that utilized pre-trained BERT embeddings in conjunction with LSTM to assess its efficacy in diagnosing TLOC (Transient Loss of Consciousness). Additionally, to investigate whether increasing the number of training epochs could enhance performance on the dataset, a comparison was made between models trained for 5 epochs and those trained for 10 epochs, as illustrated in Table 4.8.

In contrast to the first LSTM approach, there was no occurrence of classifier bias. I attribute this to the fact that BERT can comprehend the meaning of words in different contexts, while LSTM can capture long-distance dependencies.

When compared to the best-performing baseline model, the Bert+LSTM model improved Accuracy by approximately 5%. This improvement could be attributed to the additional context information extracted by the LSTM layer, leveraging BERT's characteristics. However, the context-aware LSTM model performed less effectively than the best-performing logistic regression model. It is argued that this could be due to the overfitting issue that arises when applying a relatively complex neural network to a small dataset.

As anticipated, this approach was expected to capture feature information at different levels. However, in this experiment, the combined predictions of these two models did not demonstrate significantly enhanced accuracy. And the model with more training epochs even performed worse on accuracy.There could be several potential reasons for this:

- Model Complexity vs. Data Size Mismatch: Combining complex models like BERT and LSTM demands a substantial amount of training data to effectively learn parameters. When training data is limited, models may tend to overfit, and the benefits of complex architectures may not be fully realized. Augmenting the training data with transcripts from a larger pool of patients could mitigate the risk of overfitting.

- Improper Hyperparameter Tuning: Combining two models involves adjusting additional hyperparameters, including learning rates, the number of model layers, hidden units, and more. Specific tasks require different hyperparameters, and improper tuning can prevent the models from harnessing their full potential.

- The Task Itself Doesn't Require Such Complexity: Occasionally, the complexity of text classification tasks may not necessitate the combination of two powerful models. Simple models might already suffice in addressing the problem effectively.

4.3 Experiment Three

In this experiment, we implemented speech transcription using two advanced pre-trained models, ESPnet and Whisper, on this dataset. Subsequently, we conducted an analysis of the results based on their structures and characteristics. Then, we combined the best performance ASR model and classification model, using the semantic features. The average result of four attempts with the same configuration is shown here:

4.3.1 ESPnet and Whisper

At first, we used the data to train the ESPnet model(Joint CTC-attention). Because the ESPnet does not have any constraint that guides the alignments to be monotonic as in

DNN-HMM and CTC (Watanabe et al., 2018), the attention model can be easily affected by noises. As our result shows obviously, the ESPnet indeed performs worse on the real-life data. For Whisper, there are various versions with different sizes. Three versions of five are selected from Whisper model family for the ASR task in this project. We implemented 'tiny', 'small' and 'medium' whisper models on the patient's audio.

4.3.2 Comparison of Model Performance

The comparison of the performance of these two models is shown in Table 4.9.

	Whisper-medium	Whisper-small	Whisper-tiny	ESPnet
WER	32.84%	35.39%	37.05%	53.85%
RIL	-11.25%	-10.30%	-12.77%	-12.12%
WIP	69.25%	66.89%	65.16%	46.63%
WIL	4.68%	4.53%	6.46%	21.91%

Table 4.9: ASR performance comparison

All Whisper family models outperform the ESPnet on WER metrics. WER measures the difference between the recognised text and the label text. Lower WER indicates that the model performs better in word-level recognition. As expected, the medium-sized Whisper model shows better adaptation to the patient data, with a relatively lowest WER. As lightweight models, the small and tiny-sized Whisper models also show good performance, not as good as medium, but much better than ESPnet. RIL measures how much information is lost by the ASR system during the recognition process relative to the original speech signal. Whisper-medium has the best performance on the RIL metric with -11.25%. This means that it lost 11.25% of information relative to the original speech signal. WIP measures the percentage of original information retained in the recognised text. Whisper-medium performs best on the WIP metric at 69.25%. This means that it retains about 69.25% of the original information in the generated text. Whisper family models have values above 65% on this metric, while ESPnet has only 46.63%. WIL represents the percentage of original information lost during the identification process. From the table, it can be observed that Whisper-medium and Whisper-small have the lowest WIL while ESPnet has the highest WIL.Whisper-medium and Whisper-small perform better in reducing information loss.

4.4 Experiment Four

4.4.1 Implementation of ASR + ML Classification

In this experiment, we implemented a whole pipeline for automated TLOC consultations. In this process, the Whisper-medium model, which is the best-performing ASR model in terms of metrics, is combined with the ML classification model trained with the semantic categories features. Each component is selected for the best performance and most appropriate for the patient data. The following table documents the performance of this pipeline.

Metrics	**Value**
Accuracy	0.624
Recall	0.533
F1-score	0.523

Table 4.10: Performance Metrics

For the test data, 62.4% of the results provided correct suggestions for diagnosis. Although the results of the model are not convincingly highly accurate, it still demonstrates that our model can help in diagnostic prediction.

The distribution of each classification is shown in Table 4.11.

Label/Prediction	Epilepsy	FDS	Syncope
Epilepsy	68.75%	18.75%	12.50%
FDS	31.85%	56.30%	11.85%
Syncope	25.00%	40.00%	35.00%

Table 4.11: Distribution of various types of results on the test dataset

In comparison to the previous distribution, the proportion of correctly classified cases for Epilepsy has increased by 22.75%, while the proportion of correctly classified cases for FDS has remained nearly unchanged. Syncope has also seen an increase of approximately 6% in right classifications. It can be found that the use of ASR models has an impact on the prediction of diagnosis, which improves the correct classification rate of epilepsy and syncope. However, according to the conclusions of other studies (Pevy, 2023), the use of the ASR system increases the WER, causing some information to be different from the original, therefore the model classification accuracy decreases. The reason for the differing conclusions in this experiment can be attributed to two factors. Firstly, some of the selected features may be ineffective for classification or the model has learned incorrect parameters associated with these features, resulting in more WER-transcribed texts achieving better results than manually transcribed texts. With a sufficient amount of data, the effect of randomness is reduced and we can continue to investigate the effect of each feature on the discriminatory power of the model. Secondly, the more fundamental reason is the inadequacy of the dataset, which leads to a greater susceptibility to the influence of randomness on the results.

Metrics	**Value**
Accuracy	0.776
Recall	0.673
F1-score	0.690

Table 4.12: performance metrics of Whisper-medium with ML model

After removing the Syncope labelled samples, the average results of the 5-fold cross-validation are shown in Table 4.12. 77.6% of the samples were classified into the correct category, which indicates that this method has discriminatory ability for the two seizures, epilepsy and FDS. The classification distributions are shown in the Table 4.13.

Label/Prediction	Epilepsy	FDS
Epilepsy	83.64%	16.36%
FDS	38.67%	61.33%

Table 4.13: Distribution of various types of results on the test dataset

This model shows better discrimination for epilepsy, with an accuracy of 83.64%. It is slightly weaker for FDS, with an accuracy of 61.33%.

4.4.2 Error Analysis

The models can be affected easily by the patient's accent, liaison, turbidity of words, etc. For the reason that the model was pre-trained on multi-language datasets, some words or sentences will be identified as non-English language. Another notable thing is that Whisper manages to ignore irrelevant filler sounds like "erm" and "um". It seems like the higher the model quality, the better it is at filtering out this noise. It is mainly because these pause words are not legal words in the dataset vocabulary, so the model will not identify these words though most of them are recognisable. This can be changed if we have more clinical data to fine-tune the model and let the pause words be easily identifiable to the models.

Though the Whisper model shows an acceptable level of WER on this task, there are still some typical mistakes in the ASR process. And here are some typical problems (Table 4.14).

ASR hypothesis	**Label text**	**Reason**	**WER**
YN YSTOD YCHWANEGOL ROEDDWN IN BYW YN DDIDDOROL AC YN DDIDDOROL Y DYDDOR Y DYDDOR Y DYDDOR YMWNEUD HYNNY	ER LOOKING BACK I HAD BEEN EXTREMELY BUSY AND EXTREMELY TIRED THAT WEEK ER THE WEEK RUNNING UP TO IT	Whisper model is trained on Multi-linguistic dataset, Some slurred accents lead to character mismatch	100.00%
A DWI DDIM YN OED YN YSTOD YR ATAC YN OED YNA AR HYN	I DONT REMEMBER DURING THE ATTACK ONLY BEFORE AND AFTER	Speech is fast, with elision and assimilation, leading to significant errors	100.00%

Table 4.14: Two typical mistakes and the causes in the ASR task (Whisper-medium)

Chapter 5

Conclusion

This dissertation project implemented a fully automated TLOC consultation system based on conversation analysis. The effectiveness and feasibility of two feature sets in diagnosing TLOC causes were validated. Subsequently, these feature sets were combined with traditional machine learning models to establish baseline models for TLOC cause classification. Detailed evaluations and error analyses were conducted for both binary classification and three-way classifications which incorporated speech of people with syncope.

Additionally, tokenizer and BERT embeddings were incorporated into LSTM networks for classification tasks, respectively. This endeavour aimed to bolster discernment capacity by capturing more contextual information. Subsequently, during the speech-to-text conversion process, a significant performance improvement was attained through a comparison of two pre-trained ASR models, ESPnet and Whisper, using real-world data. Besides, the experiments in this project provide detailed error analyses and classification distributions and clearly provide the effects and reflections of each method for subsequent research.

The most effective elements from each phase were integrated to establish a system capable of aiding in TLOC diagnosis through speech analysis. Promisingly, this system demonstrated impressive accuracy on the test set and holds potential for clinical applications, providing support to individuals affected by TLOC and contributing to our society.

Chapter 6

Future Work

There are more advanced models and methods to consider, such as the Transducer-based end-to-end ASR available on the ESPnet platform, the wav2vec2.0 pre-trained model as an encoder, and Streaming Transformer/Conformer ASR with blockwise synchronous beam search, among others.

However, it's important to note that these advanced models and techniques rely heavily on abundant and high-quality data. Given the challenges in collecting patient data for this project, where we currently have only a limited dataset of a few dozen patients, it is insufficient to support high-performance models. Therefore, in future diagnostic processes, it would be beneficial to collect more patient data in a more convenient and stable manner. One potential approach could involve using a dementia assessment tool like CognoSpeak[100], which utilizes a 'talking head' to ask questions and is accessible on laptops, tablets, or smartphones. This could significantly enhance the efficiency and quality of patient data acquisition.

Furthermore, due to the uniqueness and complexity of the data, fine-tuning techniques can be employed to better adapt the models to the hospital data we have collected. Fine-tuning has become a common technique in the field of computer science to improve model performance. It is highly likely that fine-tuning could enhance the model's performance in predicting TLOC causes.

Bibliography

W. O. Adkisson and D. G. Benditt. Pathophysiology of reflex syncope: a review. *Journal of Cardiovascular Electrophysiology*, 28:1088–1097, 2017. doi: 10.1111/jce.13266.

Sabur Ajibola Alim and N Khair Alang Rashid. *Some commonly used speech feature extraction algorithms*. IntechOpen London, UK:, 2018.

Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

Heather Angus-Leppan. Diagnosing epilepsy in neurology clinics: a prospective study. *Seizure*, 17(5):431–436, 2008.

Ali A Asadi-Pooya, Jennifer Tinker, and Elizabeth Fletman. Semiological classification of psychogenic nonepileptic seizures. *Epilepsy & Behavior*, 64:1–3, 2016.

Sneha Basak, Himanshi Agrawal, Shreya Jena, Shilpa Gite, Mrinal Bachute, Biswajeet Pradhan, and Mazen Assiri. Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. *CMES-Computer Modeling in Engineering & Sciences*, 135(2), 2023.

Selim R Benbadis and W Allen Hauser. An estimate of the prevalence of psychogenic non-epileptic seizures. *Seizure*, 9(4):280–281, 2000.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Michele Brignole, Angel Moya, Frederik J De Lange, Jean-Claude Deharo, Perry M Elliott, Alessandra Fanciulli, Artur Fedorowski, Raffaello Furlan, Rose Anne Kenny, Alfonso Martín, et al. Practical instructions for the 2018 esc guidelines for the diagnosis and management of syncope. *European heart journal*, 39(21):e43–e80, 2018.

P. Brignole M., Alboni. Guidelines on management (diagnosis and treatment) of syncope. *European Heart Journal*, 22:1256–1306, 2001. doi: 10.1053/euhj.2001.2739.

R. J. C. Brown and M. Reuber. Towards an integrative theory of psychogenic non-epileptic seizures (pnes). *Clinical Psychology Review*, 47:55–70, 2016a. doi: 10.1016/j.cpr.2016.06.003.

Richard J Brown and Markus Reuber. Psychological and psychiatric aspects of psychogenic non-epileptic seizures (pnes): a systematic review. *Clinical Psychology Review*, 45:157–182, 2016b.

E. Cardea, S. Pick, and R. Litwin. Differentiating psychogenic nonepileptic from epileptic seizures: a mixed-methods, content analysis study. *Epilepsy &Amp; Behavior*, 109:107121, 2020. doi: 10.1016/j.yebeh.2020.107121.

D. R. Chadwick and D. J. Smith. The misdiagnosis of epilepsy. *BMJ*, 324:495–496, 2002. doi: 10.1136/bmj.324.7336.495.

Rupali S Chavan and Ganesh S Sable. An overview of speech recognition using hmm. *International Journal of Computer Science and Mobile Computing*, 2(6):233–238, 2013.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

N Colman, K Nahm, JG Van Dijk, JB Reitsma, W Wieling, and H Kaufmann. Diagnostic value of history taking in reflex syncope. *Clinical Autonomic Research*, 14:i37–i44, 2004.

Cooper and P. R.and Westby. Synopsis of the national institute for health and clinical excellence guideline for management of transient loss of consciousness. *Annals of Internal Medicine*, 155:543, 2011. doi: 10.7326/0003-4819-155-8-201110180-00368.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

J. G. Dijk, D. G. Thijs, R. D.and Benditt, and W. Wieling. A guide to disorders causing transient loss of consciousness: Focus on syncope. *Nat Rev Neurol*, 5:438–448, 2009. doi: 10.1038/nrneurol.2009.99.

Mohit Dua, Akanksha, and Shelza Dua. Noise robust automatic speech recognition: review and analysis. *International Journal of Speech Technology*, pages 1–45, 2023.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

Kirsten M Fiest, Khara M Sauro, Samuel Wiebe, Scott B Patten, Churl-Su Kwon, Jonathan Dykeman, Tamara Pringsheim, Diane L Lorenzetti, and Nathalie Jetté. Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. *Neurology*, 88(3):296–303, 2017.

Aikaterini Fitsiori, Shivaprakash Basavanthaiah Hiremath, José Boto, Valentina Garibotto, and Maria Isabel Vargas. Morphological and advanced imaging of epilepsy: beyond the basics. *Children*, 6(3):43, 2019.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL `http://arxiv.org/abs/1606.08415`.

C Hingray, J Biberon, W El-Hage, and B De Toffol. Psychogenic non-epileptic seizures (pnes). *Revue neurologique*, 172(4-5):263–269, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997a.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997b.

Yan-Jia Huang, Yi-Ting Lin, Chen-Chung Liu, Lue-En Lee, Shu-Hui Hung, Jun-Kai Lo, and Li-Chen Fu. Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:947–956, 2022.

Gail Jefferson. *An exercise in the transcription and analysis of laughter.* Tilburg Univ., Department of Language and Literature, 1983.

Gail Jefferson. An exercise in the transcription and analysis of laughter. In Teun A. van Dijk, editor, *Handbook of Discourse Analysis*, volume 3, page 2534. Academic Press, London, 1985.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Gulbahar Karatas. Speech recognition: Everything you need to know in 2023, 2023. URL `https://research.aimultiple.com/speech-recognition/`.

Sangwook Kim, Swathi Kavuri, and Minho Lee. Deep network with support vector machines. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part I 20*, pages 458–465. Springer, 2013.

Ritsuko Kohno, Wayne O Adkisson, and David G Benditt. Tilt table testing for syncope and collapse. *Herzschrittmachertherapie und Elektrophysiologie*, 29(2), 2018.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356, 2020.

Z. Liu, L. Proctor, P. N. Collier, and X. Zhao. Automatic diagnosis and prediction of cognitive decline associated with alzheimers dementia through spontaneous speech. *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2021. doi: 10.1109/icsipa52582.2021.9576784.

E. Loper and S. Bird. Nltk: the natural language toolkit. 2002. doi: 10.48550/arxiv.cs/0205028.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Kristina Malmgren, Markus Reuber, and Richard Appleton. Differential diagnosis of epilepsy. *Oxford textbook of epilepsy and epileptic seizures*, pages 81–94, 2012.

K. Matton, M. G. McInnis, and E. M. Provost. Into the wild: transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. *Interspeech 2019*, 2019a. doi: 10.21437/interspeech.2019-2698.

Katie Matton, Melvin G McInnis, and Emily Mower Provost. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. In *Interspeech*, 2019b.

Katie Matton, Melvin G McInnis, and Emily Mower Provost. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. In *Interspeech*, 2019c.

M. R. Mehl. Quantitative text analysis. *Handbook of Multimethod Measurement in Psychology.*, pages 141–156, 2006. doi: 10.1037/11383-011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Bahman Mirheidari, Daniel Blackburn, Markus Reuber, Traci Walker, and Heidi Christensen. Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of interspeech*, pages 1220–1224. ISCA, 2016.

Pratheeksha Nair. The dummys guide to mfcc, 2018. URL `https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd`.

Mahyar Najibi, Mohammad Rastegari, and Larry S Davis. G-cnn: an iterative grid based object detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2369–2377, 2016a.

Mahyar Najibi, Mohammad Rastegari, and Larry S Davis. G-cnn: an iterative grid based object detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2369–2377, 2016b.

NICE. Epilepsies in children, young people and adults, 2023. URL `https://www.nice.org.uk/guidance/ng217`.

S. Noachtar and J. Rmi. The role of eeg in epilepsy: a critical review. *Epilepsy & Amp; Behavior*, 15:22–33, 2009. doi: 10.1016/j.yebeh.2009.02.035.

William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

N. O'Flynn. Nice guideline: Transient loss of consciousness (blackouts) in adults and young people. *Br J Gen Pract*, 61:40–42, 2011. doi: 10.3399/bjgp11x548965.

R. Pascanu, J. W. Stokes, H. Sanossian, M. Marinescu, and A. A. Thomas. Malware classification with recurrent networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015. doi: 10.1109/icassp.2015.7178304.

J. W. Pennebaker and A. Graybeal. Patterns of natural language use: disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10:90–93, 2001. doi: 10.1111/1467-8721.00123.

Nathan Pevy. Improving diagnostic procedures for epilepsy through automated recording and analysis of patients history. *White Rose eTheses Online*, 2023.

L. Plug and M. Reuber. Making the diagnosis in patients with blackouts: it's all in the history. *Practical Neurology*, 9:4–15, 2009. doi: 10.1136/jnnp.2008.161984.

Reuber M. Plug L. Making the diagnosis in patients with blackouts: It's all in the history. *Practical Neurology*, 9:4–15, 2009. doi: 10.1136/jnnp.2008.161984.

O. Press and L. Wolf. Using the output embedding to improve language models. *Proceedings of the 15th Conference of the European Chapter of The Association for Computational Linguistics: Volume 2*, 2017. doi: 10.18653/v1/e17-2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

G. H. Rawlings, I. H. Brown, B. Stone, and M. Reuber. Written accounts of living with epilepsy: a thematic analysis. *Epilepsy &Amp; Behavior*, 72:63–70, 2017a. doi: 10.1016/j.yebeh.2017.04.026.

G. H. Rawlings, I. H. Brown, B. Stone, and M. Reuber. Written accounts of living with psychogenic nonepileptic seizures: a thematic analysis. *Seizure*, 50:83–91, 2017b. doi: 10.1016/j.seizure.2017.06.006.

M. Reuber, M. Chen, J. Jamnadas-Khoda, Broadhurst, Wall M., M. M., Grnewald, and D. C. R. A., Hesdorffer. Value of patient-reported symptoms in the diagnosis of transient loss of consciousness. *Neurology*, page 10.1212/WNL.0000000000002948, 2016. doi: 10.1212/wnl.0000000000002948.

Markus Reuber, Guillen Fernandez, Jürgen Bauer, Christophe Helmstaedter, and Christian E Elger. Diagnostic delay in psychogenic nonepileptic seizures. *Neurology*, 58(3):493–495, 2002.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2 (2):1–26, 2018.

Ploypaphat Saltz, Shih Yin Lin, Sunny Chieh Cheng, and Dong Si. Dementia detection using transformer-based deep learning and natural language processing models. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 509–510. IEEE, 2021.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Meike Schwabe, Stephen J Howell, and Markus Reuber. Differential diagnosis of seizure disorders: a conversation analytic approach. *Social science & medicine*, 65(4):712–724, 2007.

Meike Schwabe, Markus Reuber, Martin Schondienst, and Elisabeth Gulich. Listening to people with seizures: how can linguistic analysis help in the differential diagnosis of seizure disorders? *Communication & medicine*, 5(1):59, 2008.

S. J. M. Smith. Eeg in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery amp; Psychiatry*, 76:ii2–ii7, 2005. doi: 10.1136/jnnp.2005.069245.

K. J. Staley. Molecular mechanisms of epilepsy. *Nature Neuroscience*, 18:367–372, 2015. doi: 10.1038/nn.3947.

I. G. Stiell and C. Bennett. Implementation of clinical decision rules in the emergency department. *Academic Emergency Medicine*, 14:955–959, 2007. doi: 10.1197/j.aem.2007.06.039.

Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. Automated recognition of alzheimers dementia using bag-of-deep-features and model ensembling. *IEEE Access*, 9:88377–88390, 2021.

Martijn R Tannemaat, Julius van Niekerk, Robert H Reijntjes, Roland D Thijs, Richard Sutton, and J Gert van Dijk. The semiology of tilt-induced psychogenic pseudosyncope. *Neurology*, 81(8):752–758, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017a.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Dong Wang, Xiaodong Wang, and Shaohe Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018, 2019.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878. IEEE, 2020.

A. Wardrope, E. Newberry, and M. Reuber. Diagnostic criteria to aid the differential diagnosis of patients presenting with transient loss of consciousness: A systematic review. *Seizure*, 61:139–148, 2018. doi: 10.1016/j.seizure.2018.08.012.

A. Wardrope, J. Jamnadas-Khoda, R. A. Grnewald, T. J. Heaton, S. Howell, and M. Koepp, M. J. Reuber. Machine learning as a diagnostic decision aid for patients with transient loss of consciousness. *Neurol Clin Pract*, 10:96–105, 2019. doi: 10.1212/cpj.0000000000000726.

Alistair Wardrope, Jenny Jamnadas-Khoda, Mark Broadhurst, Richard A Grünewald, Timothy J Heaton, Stephen J Howell, Matthias Koepp, Steve W Parry, Sanjay Sisodiya, Matthew C Walker, et al. Machine learning as a diagnostic decision aid for patients with transient loss of consciousness. *Neurology: Clinical Practice*, 10(2):96–105, 2020.

S. Watanabe, T. Hori, S. Kim, J. W. Hershey, and T. Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11:1240–1253, 2017. doi: 10.1109/jstsp.2017.2763455.

S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Yalta, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. Espnet: end-to-end speech processing toolkit. *Interspeech 2018*, 2018. doi: 10.21437/interspeech.2018-1456.

A. J. Whitfield, S. Walsh, and M. Reuber. Catastrophising and repetitive negative thinking tendencies in patients with psychogenic non-epileptic seizures or epilepsy. *Seizure*, 83: 57–62, 2020. doi: 10.1016/j.seizure.2020.09.034.

Ying Xu, Dennis Nguyen, Armin Mohamed, Cheryl Carcel, Qiang Li, Mansur A Kutlubaev, Craig S Anderson, and Maree L Hackett. Frequency of a false positive diagnosis of epilepsy: a systematic review of observational studies. *Seizure*, 41:167–174, 2016.

Xue Lanny Y and Anthony L Ritaccio. Reflex seizures and reflex epilepsy. american journal of electroneurodiagnostic technology. *Clinical Autonomic Research*, 14:i37–i44, 2004. doi: 10.1007/s10286-004-1006-0.

Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. Deep learning-based speech analysis for alzheimers disease detection: A literature review. *Alzheimer's Research & Therapy*, 14(1):1–16, 2022.

Amir Zaidi, Peter Clough, Paul Cooper, Bruce Scheepers, and Adam P Fitzpatrick. Misdiagnosis of epilepsy: many seizure-like attacks have a cardiovascular cause. *Journal of the American College of Cardiology*, 36(1):181–184, 2000.

J. Zentner, H. K. Wolf, Helmstaedter, Grunwald C., Aliashkevich T., Wiestler A. F., and J. O. D. Schramm. Clinical relevance of amygdala sclerosis in temporal lobe epilepsy. *Journal of Neurosurgery*, 91:59–67, 1999. doi: 10.3171/jns.1999.91.1.0059.